

In defence of DARPA

The Pentagon's boldest research agency is in trouble; its unique character is in real danger. DARPA's officials and those who have benefited from its largesse must engage in open debate about the agency's value.

If you plan to visit the US Defense Advanced Research Projects Agency (DARPA), be sure to take a map. The agency resides in a nondescript, black-panelled building in a Washington suburb; its only distinguishing feature is a police car parked in front of the main entrance 24 hours a day.

Like the building it calls home, DARPA can be aptly described as a black box. Its funding decisions are determined not by peer review, but by the whims of programme managers, usually seconded from academia or industry. Their sole guidance is DARPA's mission of backing imaginative, creative and high-risk research in the interest of the defence of the United States. Scientists who capture the imagination of a programme manager can gain support for projects that would be dismissed by other agencies as too speculative to win public funding.

This freewheeling approach has yielded some big payoffs, such as the radar-absorbing skin of 'stealth' aircraft, and a host of critical Internet technologies. Today, the agency is pioneering the development of spintronics and quantum computing, which together could transform information technology. It has also extended its innovative reach into systems neuroscience, and other areas of biology.

But DARPA's capricious funding methods have led it down some strange roads. In the early 1970s, the agency dispatched a team to test the psychic prowess of spoon-bender Uri Geller. More recently, DARPA devoted hundreds of thousands of dollars to disputed research into tabletop nuclear 'bubble fusion'.

But earlier controversies are nothing compared with the storm whipped up by DARPA's Information Awareness Office. Created to investigate new intelligence-gathering techniques in the wake of the terrorist attacks of 11 September 2001, the office was headed by John Poindexter — who, as President Reagan's national security adviser in the 1980s, was a central figure in the Iran-Contra scandal.

Poindexter's office has trampled close to cultural taboos, especially civil liberties — an early programme to monitor financial records for terrorist activity contained disturbing echoes of 'Big Brother'. Now a study that sought to predict threats by asking investors to bet on a 'futures market' in terrorism, known as FutureMAP, has triggered an intense storm. Poindexter has tendered his resignation, and DARPA's congressional overseers are on the warpath (see page 601).

By its very nature, DARPA funds projects that critics will be able to depict as absurd or sinister. If the agency is to hold up under the scrutiny that it can now expect, it must overhaul its 'black box' image. Historically, DARPA has given little thought to public and media relations — and recent actions have shown it to be naive in this sphere. At the height of the terrorist-futures controversy, the project's web page disappeared, and this year a misguided Information Awareness Office logo sporting an unblinking eye staring down on the world was similarly axed. Neither really disappeared: privacy groups quickly posted the vanished pages on their own websites; G-strings and coffee mugs sporting the old logo can be purchased online. In the end, these moves only made DARPA look like it had something to hide.

Rather than trying to sweep its mistakes under the carpet, DARPA should be open and clear about the research it funds. It must engage in public debate about its more controversial work, and explain why pushing back the boundaries of our understanding requires occasional forays into areas that sound absurd or ethically questionable.

Meanwhile, scientists who have benefited from DARPA's unique approach should remind politicians and the public of what would be lost were the agency to be reined in. DARPA has existed since 1958 as a one-of-a-kind research organization. It is unlikely to disappear. But if its supporters fail to speak out, the ingenuity that distinguishes DARPA from the pack very well could. ■

Hubble versus the future

NASA's science managers should be cut some financial slack to prevent the sacrifice of a prized research asset.

What could be sillier than throwing away the Hubble Space Telescope, one of the most productive instruments in the history of science? That was the message from most speakers at last week's public meeting on the telescope's fate (see page 603). Yet there was dissent, mostly from people whose projects face delays if Hubble gains an extended life beyond its planned demise in 2010.

Foremost among those projects is the James Webb Space Telescope (JWST). Despite being conceived as the Next Generation Space Telescope, it is not Hubble's successor. It will observe in the near- and mid-infrared, where Hubble is not so competitive. In a decadal review of priorities published in 2000, US astronomers asked for a large infrared space telescope, not for Hubble 2 — and that's just what NASA is giving them. But circumstances have changed. The JWST was due to launch in 2007, but now will not arrive until 2011 at the earliest. Nervous astronomers have come to see the value of a bird in the hand.

At last week's meeting, Bruce McCandless, a former astronaut who helped to place Hubble in orbit in 1990, suggested that major

instruments should not be decommissioned until user demand drops. Other satellites, such as the International Ultraviolet Explorer and the Compton Gamma Ray Observatory, were arguably past their prime when they were shut down. But observing time on Hubble looks set to remain as coveted as ever for many years to come.

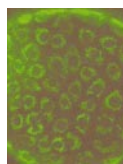
Unfortunately, NASA can't pay for the JWST and other future projects without retiring Hubble, unless Congress and the White House decide to loosen the purse strings. NASA's science managers have already been rewarded with budget increases for their relative financial rectitude, while those in charge of the space station have been punished for their profligacy. Astronomers can make a strong case for further cash injections into NASA's science programme, to delay Hubble's demise without sacrificing the JWST's timetable.

In today's harsh economic climate, there's no guarantee of success. But that shouldn't stop NASA earmarking small sums for preliminary studies of future Hubble instruments, to preserve the option of upgrading the telescope again after 2005 if the cash is forthcoming. ■





Hubble trouble
NASA contemplates
the demise of its
space telescope
p603



Matter transfer
Could sperm offer
a cheap route to
transgenic animals?
p604



Paper jam
Europe's science
funding system leaves
researchers fuming
p605



Cash cows
Australia secures
stake in bovine
genome sequence
p606

Terrorist betting leaves defence agency fighting for autonomy

Geoff Brumfiel, Washington

Chemical attack or assassination — which would you bet on as most likely to happen?

The question is now academic, following the cancellation of a US defence project to run an online futures market on terrorist acts. But the fallout from the scheme is threatening to cost the Defense Advanced Research Projects Agency (DARPA) its prized independence.

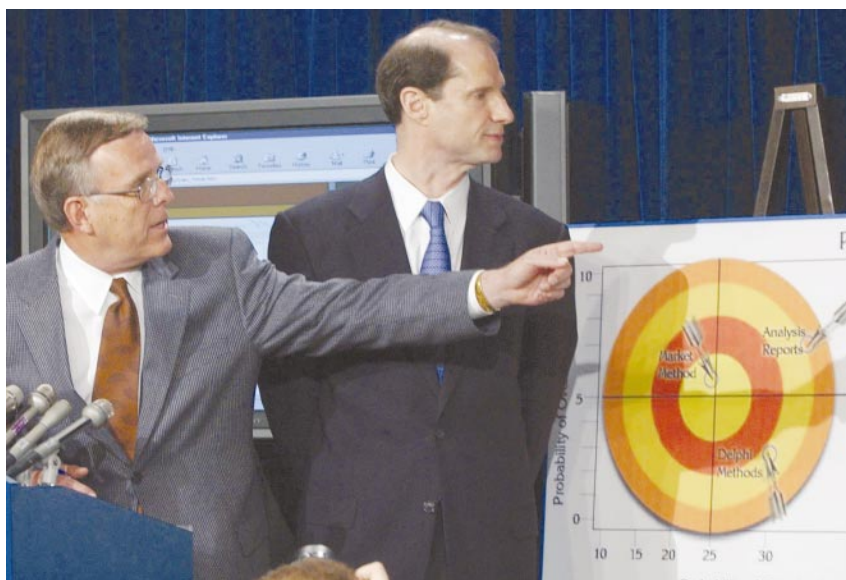
The project, known as FutureMAP (Futures Markets Applied to Prediction), was set up by DARPA to see whether trading on a futures market could help to predict the likelihood of a terrorist act being committed. Investors would have been able to bet on questions such as: "Will terrorists attack Israel with bioweapons next year?"

But the scheme sparked a public outcry and on 29 July, defence secretary Donald Rumsfeld pulled the plug on it. Now law-makers and the Pentagon's top brass are considering whether DARPA's independent approach to research should be reined in.

DARPA, the research arm of the Department of Defense, spends more than US\$3 billion each year on unconventional research ranging from quantum computing to biomimetics. Historically, it has enjoyed relative freedom from the defence department's bureaucracy, and its innovation is widely acknowledged. Clipping its wings "would be a disaster for the country", warns James Harris, an electrical engineer at Stanford University in California.

Law-makers blasted DARPA for funding FutureMAP — Senators Byron Dorgan (Democrat, North Dakota) and Ron Wyden (Democrat, Oregon) branded the scheme as unethical at a press conference on 28 July. Republicans joined the Democrats' call for a swift end to the research, and within two days the programme was terminated, and the man responsible for it, John Poindexter, had offered his resignation. Poindexter, who is director of DARPA's Information Awareness Office, is no stranger to controversy — he was President Ronald Reagan's national security adviser during the Iran-Contra scandal.

This is not the first time that DARPA has been lambasted for poor taste over one of



All bets are off: Byron Dorgan (left) and Ron Wyden lambast DARPA's terrorist futures market.

Poindexter's projects. Last autumn, the Total Information Awareness programme to search for patterns in electronic databases of public and private records was heavily criticized by civil libertarians and the media. The Senate axed funding for that programme earlier last month.

But not all of DARPA's programmes are problematic. Many credit the agency with developing the Internet, and it has long been a favourite among bright researchers with unconventional ideas. "Compared with agencies such as the National Science Foundation, DARPA takes much bigger chances," says Harris, who has DARPA funding to develop a thumbnail-sized fluorometer that can detect chemical and biological agents. DARPA's generous funding and goal-oriented programmes allow researchers to "do more than just incremental science", adds Charles Lieber, a nanotechnology researcher at Harvard University.

At times the agency's free-thinking programmes have pushed technology ahead, despite initial resistance from authorities. In the late 1970s, for example, DARPA researchers began work on unmanned aerial vehicles

against the wishes of the US Air Force. When the first unmanned vehicle was ready for testing in 1998, "the undersecretary of defence forced the Air Force to accept it", recalls Frank Fernandez, an engineer at the Stevens Institute of Technology in Hoboken New Jersey, who headed DARPA from 1998 to 2001.

But the public questions of ethics and civil liberties surrounding the current projects are generating far more controversy than previous debates, comments former DARPA project manager Randy Katz, a computer scientist at the University of California, Berkeley.

Members of Congress are now calling for a closer look at DARPA's programmes. The Senate committee overseeing the agency is pushing for a new external panel, appointed by the defence secretary, to help set DARPA's research agenda. "This latest episode is very damaging," says one senate staff member. DARPA director Tony Tether did not respond to *Nature's* request for an interview.

The fresh scrutiny worries many advocates of the agency's occasionally fringe research programmes. "If they start clamping down on DARPA, they will kill it," warns Fernandez.

Fast vaccine offers hope in battle with Ebola

Tom Clarke and Jonathan Knight

High-speed vaccines could help to contain unexpected outbreaks of disease, say US researchers who have developed a fast-acting vaccine against the deadly Ebola virus.

The new vaccine, reported in this issue of *Nature* (see page 681), gives macaque monkeys protection from Ebola infection only four weeks after a single jab. Previous Ebola vaccines took six months and multiple boosters to confer immunity.

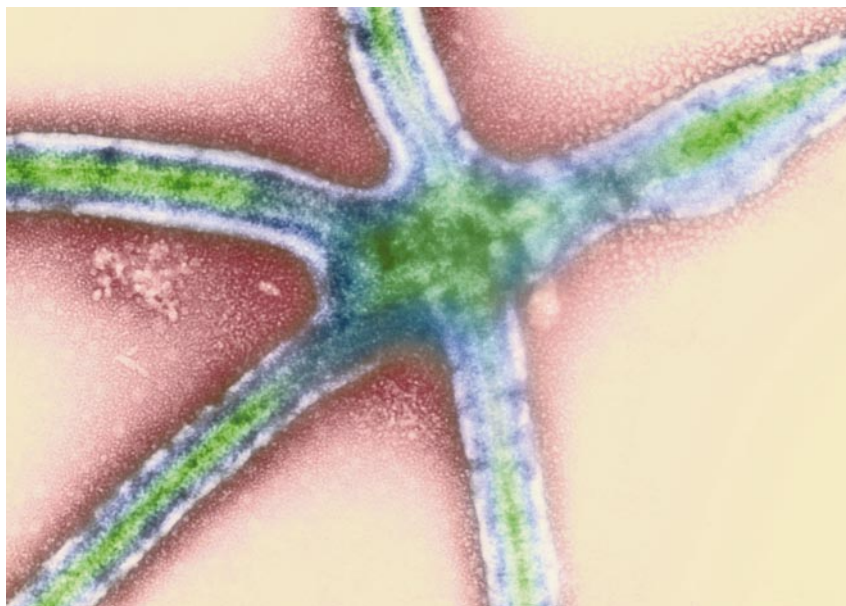
Should the vaccine prove safe and effective in humans, this would be fast enough to stop Ebola outbreaks, says lead author Gary Nabel, head of the Vaccine Research Center at the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland. It could also form a template for developing emergency vaccines against unexpected outbreaks of disease, Nabel says.

Most vaccines serve only as a preventive measure, because they take too long to produce immunity to be useful once an outbreak has begun. A notable exception is smallpox vaccine, which acts so quickly that it can sometimes stop the disease even after a person has been infected with the virus.

Fast vaccines permit a strategy called ring vaccination, which aims to contain an outbreak by inoculating all possible contacts of the first detected cases — and contacts of contacts as well. Ring vaccination is considered one viable option for dealing with a bioterror attack using smallpox.

Nabel believes that it would also be effective against Ebola outbreaks, which occur every few years in parts of Africa. Ebola virus causes a severe and untreatable haemorrhagic fever that kills up to 90% of its victims by liquefying their organs.

To make the new vaccine, Nabel's team stitched a portion of the Ebola virus genome into a modified adenovirus, a cause of the common cold. Adenovirus generates a very aggressive immune response in the host, and so teaches the immune system to recognize



In the pink: researchers hope their new vaccine against Ebola (pictured) will help to contain outbreaks.

the protein encoded by the inserted gene.

Macaques injected with the vaccine became immune to Ebola within 28 days. Infectious-disease experts expressed surprise over news of the vaccine's speed. "My reaction is 'wow'," says Steve Wolinsky, who heads the infectious-disease division at Northwestern University in Chicago, Illinois. If this could be reproduced with other pathogens, it would dramatically help the control of viral disease, he says.

One trade-off in producing a fast interventionist vaccine is that the one-shot approach gives a weaker immune response than conventional vaccination with multiple jabs. But this may not matter, as the response was sufficient to prevent disease. "During an Ebola outbreak that's all you need," says virologist Yoshihiro Kawaoka at the University of Tokyo.

What remains to be determined is how long immunity lasts. Ebola outbreaks can last

up to a year and the vaccine would need to confer protection for that long, Kawaoka warns.

There are other hurdles too. About 45% of the US population carries protective antibodies against adenovirus and might be resistant to the vaccine. Nabel's group is working on ways to coat the adenovirus particles chemically to shield them from rejection by the immune system.

The possibility of fast-acting vaccines should nevertheless prompt a rethink of vaccination strategy at times when rapid responses are needed to viral bioweapons or novel viruses such as that which causes severe acute respiratory syndrome (SARS), says Nabel. "This could change our thinking about how we use vaccines," he says.

Nabel's team is collaborating with the biotechnology company GenVec in Gaithersburg, Maryland, to produce a SARS vaccine using the adenovirus technology. ■

Czech stem-cell work heightens calls for EU ruling

Alison Abbott

Czech scientists say they have derived three human embryonic stem-cell lines from spare embryos stored at an *in vitro* fertilization clinic in Brno.

This makes the Czech Republic the first of the eastern European countries poised to join the European Union (EU) to move into this controversial research area.

It also adds to pressure on the EU to decide whether to fund research on newly derived stem-cell lines. This research is

allowed under strict ethical supervision in some EU countries, such as Britain and Sweden, but is banned in others, including Germany and Italy. Last week the Spanish government changed sides and approved a proposed law to allow the production of cell lines from spare embryos for research.

The Czech Republic has no law controlling human embryonic stem-cell research. But Eva Syková, head of the Centre for Cell Therapy and Tissue Repair at Charles University in Prague, who developed the three cell lines

together with colleagues at the Mendel University of Agriculture and Forestry in Brno, says they are working to high ethical standards. They received informed consent from donor couples undergoing *in vitro* fertilization, she points out.

The scientists are now characterizing the lines, and plan to study the cells' potential to develop into differentiated cells such as neurons, which they believe could have therapeutic potential. They presented their results at a meeting in Prague last month. ■

NASA under pressure to extend Hubble's life

Tony Reichhardt, Washington

More than 200 leading astronomers crowded into a meeting room in Washington DC last week to grapple with a thorny dilemma: when, or even whether, to shut down the 13-year-old Hubble Space Telescope.

NASA wants to decommission Hubble in 2010, saving the money to prepare the larger and more capable James Webb Space Telescope (JWST) for launch in 2011 (see *Nature* **416**, 112; 2002). But a growing number of astronomers think that would be a mistake. Riccardo Giacconi, a Nobel laureate and former director of the Space Telescope Science Institute in Baltimore, Maryland, warned that they "would look like fools" if they didn't keep the highly productive Hubble going at least until the JWST is launched.

So charged has the debate become that NASA has thrown the question over to a heavyweight panel of astronomers led by John Bahcall of the Institute for Advanced Study in Princeton, New Jersey. The panel will solicit views from astronomers and report to NASA in October.

NASA currently favours a last upgrade for Hubble in 2005, before attaching a rocket booster to the telescope in 2010 that would steer it to burn up over the ocean. A team at the agency's Marshall Space Flight Center in Huntsville, Alabama, is studying how to do this, with a report due at the end of this month. Many astronomers say it would add relatively little to the \$700-million cost of the 2010 trip to prolong Hubble's life, and even

add new instruments. The Marshall team considers this to be too difficult at present.

In an extraordinary gesture of support for Hubble, astronaut John Grunsfeld said that US astronauts opposed "risking human lives for the purpose of disabling great science", but would support a mission to extend Hubble's life or ensure its safe re-entry.

But George Rieke, an astronomer at the University of Arizona who is building one of the JWST's instruments, said that it would be unwise to improve Hubble's ability to survey distant galaxies in the infrared when "we're building another instrument to do that even better".

Giacconi, however, pointed to Hubble's

unique ability to see across visible, ultraviolet and infrared wavelengths — something that the more sensitive JWST won't be able to match. Astronomers viewing in X-rays and γ -rays have come to rely on supporting observations from Hubble, he said.

John Huchra of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, who chairs the Association of Universities for Research in Astronomy (AURA), called Hubble "the Energizer Bunny of astronomy". He said that a poll of AURA members ranked continued space-based observations in visible and ultraviolet wavelengths as their highest priority. ■

♦ <http://hst-jwst-transition.hq.nasa.gov/hst-jwst>



Shine on: many astronomers want the Hubble Space Telescope to remain in service after 2010.

Ecological advice sparks sea change in judicial opinion

Rex Dalton

For any environmental scientist, it would be an opportunity to die for — to speak to the judges of one of the most powerful US courts about the ecological perils to the oceans.

On 25 June, this dream became a reality for marine ecologists Jeremy Jackson and Nancy Knowlton from the Scripps Institution of Oceanography in San Diego, California.

They were invited to provide scientific background to the US Ninth Circuit Court, the largest federal circuit court network and one of the country's most powerful courts. Some 1,600 judges and attorneys attended its annual conference in Kauai, Hawaii.

Cases covering marine-mammal protection, fishing standards and disputes over seismology or sonar studies are often heard in the Ninth Circuit Court, which sets precedents for US waters off the West Coast, Hawaii and Alaska.

The unusual idea of inviting the



scientists came from Judge Raymond Fisher, a member of the Ninth Circuit Court of Appeals. The judges and attorneys "were interested in hearing the scientific side of things directly from scientists, rather than filtered through court proceedings," says

Fisher. The audience was treated to an unvarnished look at marine ecological issues — and it made them think differently.

"The ocean is really in very bad shape," Jackson told the audience, citing the collapse of major fisheries, degradation of coral-reef habitats and the widespread growth of slime from oxygen depletion and agricultural chemical pollution.

"Afterwards, people in the audience said they were largely unaware of the seriousness of the situation," says Knowlton. "They said they may want to rethink how they address certain environmental issues."

William Fletcher, a Ninth Circuit appellate judge who helped to select the scientists, added: "People thought the talks were fabulous. I thought I knew a lot about the environment, but I was staggered by what I didn't know."

Court officials say they expect to invite scientists to next year's conference. ■

Piglets add some colour to transgenic story

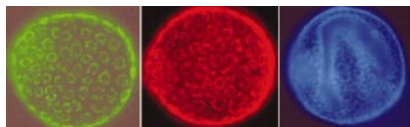
Nicola Nosengo

Researchers in Italy and Australia are to introduce 18 transgenic piglets to a meeting in California next week, in an attempt to convince sceptics that their disputed transgenic methodology does indeed work. They say that the technique could be a cheap and efficient way to produce animals expressing multiple foreign genes.

Marialuisa Lavitrano of the University of Milan-Bicocca, together with a group led by Ian McKenzie at the Austin Research Institute in Melbourne, created the animals, which express up to three foreign test genes, using a method called sperm-mediated gene transfer (SMGT).

Lavitrano first reported the method in 1989, when she found that sperm cells can internalize foreign DNA fragments, integrate the fragments' genes into their chromosomes, and then transfer these genes to the next generation during fertilization. But many laboratories were unable to repeat her results, probably because key factors — such as optimal incubation temperature and the time required for DNA integration — vary widely, even between animals of the same species.

Lavitrano hopes to win over the doubters when she discusses the piglets on 11 August



Colour code: a pig embryo, seen through colour filters, reveals expression of three foreign genes.

at the Transgenic Animal Research Conference in Lake Tahoe, California. Together with her Australian colleagues, she transferred test genes encoding coloured proteins — green, red and blue — into sperm cells and used them to inseminate female pigs.

The technique was surprisingly successful: of more than 100 embryos created, 90% expressed all three genes. And when two of the sows gave birth in Bologna, Italy, on 20 April, seven piglets contained all three coloured proteins, another seven contained two, and the remaining four displayed one.

The results are “big news”, says Bob Wall, a researcher with the US Department of Agriculture in Beltsville, Maryland. SMGT — in the hands of Lavitrano and McKenzie, at least — has a better success rate than other, more widely used transgenic techniques, such as DNA micro-injection and nuclear transfer.

DNA micro-injection, which involves injecting DNA fragments into a fertilized egg, results in fewer than 1% of injected eggs giving rise to healthy animals expressing transgenes. And nuclear transfer, in which the nucleus of a genetically engineered adult cell is transferred to an egg cell whose nucleus has been removed, has an efficiency of 1–2% in livestock, and has not yet led to successful multiple gene transfer.

Animals expressing multiple foreign genes will be important in studying the feasibility of organ transplantation from pigs to people, says Lavitrano. Introducing a suite of genes to make pig organs more similar to those of humans could help to prevent their rejection by the recipient's immune system, she says.

But companies still doubt that SMGT will be commercially viable. Tom Newberry of Genzyme Transgenics Corporation in Framingham, Massachusetts, says that micro-injection and nuclear transfer are inefficient but more commercially attractive, as they do not require specific lab conditions for each animal.

“SMGT could be important for smaller laboratories that cannot afford to handle the large numbers of animals that inefficient methods require,” he adds. ■

VETERINARY SCIENCE FACULTY, UNIV. BOLOGNA

WHO prepares for final push to rid the world of polio

Declan Butler

With the fight against polio close to being won, the newly appointed director-general of the World Health Organization (WHO), Lee Jong-wook, announced on 29 July that eradication of the disease by 2005 is one of his top priorities. But he said that more funds are needed to do the job properly.

Health workers agree that a final big push is needed. Poliovirus samples remain unmonitored in non-secure freezers across the world, and some vaccination programmes are lax. Any complacency threatens to undo years of gains, they say.

The WHO's US\$3-billion Global Polio Eradication Initiative, launched in 1988, has been highly successful, slashing the number of polio cases from 350,000 in 125 countries to just 235 in seven countries this year.

“But there is a false sense of security that the end is in sight,” says David Heymann, the official appointed by Lee to lead the new push. The virus must be eliminated in the few countries where it remains, warns Heymann, as even a single case could spawn an outbreak that could be exported.

Lee has launched a three-month mass vaccination campaign, involving 175 million



Closing in: a worker at Ethiopia's National Polio Laboratory in Addis Ababa.

children in India, Nigeria, Pakistan and Egypt, to eliminate polio transmission by the end of 2004. But before polio can be declared the second disease to be eradicated (after smallpox in the 1970s), sources of potential new outbreaks must be contained.

A major source might be the widely used oral polio vaccine, an attenuated live virus that can mutate back to wild, virulent poliovirus in unvaccinated populations and cause

outbreaks (see *Nature* 409, 278–280; 2001).

“If this happens often we could be in trouble,” admits Heymann. An alternative is the injectable Salk inactivated polio vaccine, which cannot cause disease. But this is hard to use efficiently in developing countries.

Poliovirus stocks held in laboratories across the world present another potential source of outbreaks. To stop this, the WHO this month launched a “Global action plan for laboratory containment of wild polioviruses”, akin to the 1970s programme that targeted smallpox stocks. This requires countries to search for stocks and either destroy them or, if they are needed to provide live virus for vaccine manufacture, store them in biosecure facilities.

There are also other potential sources of infection. Diagnostic stool samples from polio patients contain live poliovirus and are widely stored, “possibly in their thousands”, says Heymann. The WHO has responded by launching a stool inventory. Poliovirus contamination of laboratory cultures of other viruses has also been reported (M. Davies *et al.* *Lancet* 361, 1187–1188; 2003). ■

WHO

Red tape frustrates Europe's fund-seekers

Ralf Jox, Munich

Scientists who burnt the midnight oil for months to apply for the first round of funding under the European Union's new Framework research programme are angry about the complex bureaucracy they were forced to endure.

Of more than 10,000 applications in the first round of the €17.5-billion (US\$19.5-billion) Sixth Framework Programme (FP6), about a quarter have been recommended for funding, a European Commission official told *Nature* last week. This is higher than for some key areas in the equivalent stage of the Fifth Framework Programme (see *Nature* 404, 695; 2000).

But it is still below the target for most national research funding agencies, which aim to fund between 30% and 40% of applications. And in some research areas, such as cancer, the FP6 success rate is as low as 15%.

Researchers are frustrated by the cumbersome application procedure. "Application for European Union (EU) research grants is ridiculously complicated, given the relatively modest amount of money that you can expect at the end of the day," says Karl Tryg-vason, a cell biologist at the Karolinska Institute in Stockholm, Sweden, whose application has been recommended for funding.

The complexity has its roots in a new EU philosophy. This focuses funding on "solving problems that are relevant to European communities and industries" and aims to bring industry and the ten accession countries — those who will join the EU next year — into the research fold.

Most of the money is allocated through two new schemes — Integrated Projects and Networks of Excellence — which are intended to foster the formation of large consortia.



Long nights: the application process for European funding has proved complex and time consuming.

The Networks of Excellence scheme requires an interdisciplinary consortium of scientific groups from at least three countries, one of which is preferably an accession country.

The requirements are so complex that the commission urged researchers to use professional consultancies to help them prepare their applications. It admits that it received fewer applications under this scheme than it was expecting — and that of the thousand or so it did receive, many were rejected for not adhering closely enough to the requirements.

Many applications involved consortia of up to 40 groups, but this failed to impress Achilles Mitsos, director-general of the research commission. "The objectives of Networks of Excellence were not adequately understood," he says. "The main intention

was to promote integration, not sheer size."

The commission is also disappointed at the failure of some of the political aims of FP6, given the low participation of industry and the accession countries. A researcher at Danone, a Paris-based food company, is not surprised: "The Framework Programme is complex — and business needs to move fast."

The EU's aspiration to address political issues through its research programme has long been a bugbear for many scientists, who argue that it is hard to reconcile with selecting the best scientists for funding.

"Scientific excellence was at the very centre of the evaluation," says Mitsos. "But we also had to make sure that European citizens will truly benefit from European research." The commission plans to provide clearer guidelines for future calls. ■

Companies vie to put all your genes on a chip

Carina Dennis

The race is on to produce a commercial DNA microarray that covers the whole human genome.

Applied Biosystems of Foster City, California, announced on 22 July that it plans to have a 'whole-genome chip' on the market by the end of the year. In response, Affymetrix, based at Santa Clara, California, and Agilent Technologies in Palo Alto, California, immediately announced that they intend to do the same.

DNA microarrays allow researchers to assess the level of expression of thousands of genes at a time (see page 610), and a single array for the whole human genome would be an important tool, experts say. Scientists

will be queuing up to buy the arrays, says John Hogenesch, a molecular biologist at the Genomics Institute of the Novartis Research Foundation in San Diego.

No one knows how many genes there are in the human genome (see *Nature* 423, 576; 2003), but the number represented on the arrays under development ranges from 31,000 to 39,000. "I suspect that the companies are making an informed guess and early versions will be a first step — though a great first step," says Richard Young, a geneticist at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts.

The three firms are already jockeying for position. Stephen Fodor, chief executive of

Affymetrix, points out that his company has been selling custom versions of a single whole-genome array for the human genome to individual clients for several years. Agilent, meanwhile, says that it already has prototypes for its whole-genome chip.

Researchers have more to look forward to from the chip manufacturers. Both Applied Biosystems and Agilent, for example, have plans to make single whole-genome arrays for mouse and rat available next year.

And Affymetrix has developed prototype arrays for the human 'transcriptome', which represents not only each gene in the genome but also all of the variants of individual genes. ■

Spain in the frame for Europe's bid to host fusion reactor

London A committee asked to assess two proposed European sites for ITER, a US\$5-billion fusion reactor, has chosen Vandellòs in Spain above Cadarache in France.

The partners working on the research reactor — the European Union, Canada, Japan, Russia, China and the United States — hope to select a site by the end of the year. Europe believes it will have a better chance of fighting off competition from Canada and Japan, who are also preparing bids, if it puts forward a single site.

As neither France nor Spain is willing to step down, the European Commission (EC) asked an independent group of science administrators in May to examine the sites at Vandellòs, near Barcelona, and Cadarache, an existing nuclear research facility in southern France. The committee, led by David King, science adviser to the UK government, has not yet released its report, but is thought to back the Spanish site as the cheaper option.

The EC will use the report next month to help prepare a proposal for the decision-making European Council of Ministers at the end of September. France is expected to lobby against the committee's verdict.

Journal editor quits hot seat over global-warming study

Munich Three editors of the journal *Climate Research* resigned last week over a published paper claiming that current global warming falls within the range of natural variability.

The finding was seized upon by some conservative US politicians, but has been attacked by climate researchers. Critics questioned the methods used by the authors — astronomers Willie Soon and Sallie Baliunas of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts — to compare proxies for climate change such as tree-ring growth (W. Soon and S. Baliunas *Clim. Res.* 23, 89–110; 2003).

Hans von Storch, a coastal researcher at the GKSS, an environmental research centre in Geesthacht, Germany, and the journal's editor in chief, wanted to publish an editorial conceding that the paper's errors should have been spotted during peer review. Von Storch and two other editors resigned after the editorial board's remaining members failed to agree on the text of the editorial.

Otto Kinne, the journal's publisher, says he backed the editorial but that von Storch resigned before the board had given it the green light. The journal will publish a different online editorial on 5 August, accepting that it should have "insisted on solid evidence and cautious formulations before publication".

Australia beefs up an udder genome project

Sydney The CSIRO, Australia's main national research agency, last week announced plans to donate A\$1.5 million (US\$1 million) to the international effort to sequence the cow genome. The organization says that it hopes to milk the genome project for all it's worth to give Australian industry an edge over international competitors.

The US National Human Genome Research Institute in Bethesda, Maryland, provisionally approved the Bovine Genome Project in March, pledging half of the project's US\$50-million cost if the remainder can be raised from elsewhere.

Assuming that the funds can be obtained, sequencing should begin in September, mainly at Baylor College of Medicine's Human Genome Sequencing Center in Houston, Texas.



WWW.GENOME.GOV

All change at the top for French research

Paris French scientists will return from their long summer holidays at the end of this month to find a new set of leaders running their country's research effort.

In a large reshuffle, science minister Claudie Haigneré last week removed Geneviève Berger as director general of the CNRS, the national research agency. Berger, who is three years into a four-year term, will be replaced by Bernard Larrousurou, current head of the INRIA, the national computing agency. Haigneré said that Larrousurou will bring stronger leadership and a new impetus to the 25,000-strong CNRS.

Bernard Bigot, head of Haigneré's office, will become high commissioner of the Atomic Energy Commission. His post will be filled by Philippe Braidy, financial director of the national space agency, the CNES.

The reshuffle coincides with a promise by the government that research will be spared in next year's budget, and suggests that Haigneré is preparing for a busy time this autumn, when her new team starts work.

Worldwide agricultural effort aims to help world's poor

London An ambitious attempt to assess how agricultural technology can aid farming in developing countries was approved last week.

Delegates from the United Nations, industry, environmental groups and consumer organizations met from 31 July to 2 August in Budapest, Hungary, to finalize plans for the assessment, which was initiated last year by the World Bank. Over the next three years, hundreds of experts will work with local and regional stakeholders to pool knowledge about agricultural technology. The resulting compendium will provide the inhabitants of developing countries with solutions to water management, irrigation, soil erosion and other problems.

"This is the first global intergovernmental attempt to assess how agriculture can promote development, and what science and technology have to offer," says Louise Fresco, assistant director general of the United Nations' Food and Agriculture Organization.

The effort will be chaired by Bob Watson, the World Bank's chief scientist. The Intergovernmental Panel on Climate Change, of which Watson is a former director, will serve as a blueprint for the exercise.

♦ www.agassessment.org

Meeting agrees on unified vision of Earth

Washington The need for a global observation system was acknowledged last week by delegates from more than 30 nations at the Earth Observation Summit in Washington DC (see *Nature* 424, 357; 2003).

Attendees agreed to begin work on a ten-year implementation plan for the system, which will involve sharing data from hundreds of sensors on the ground, in the ocean and in space. The scheme will also increase aid to developing countries to help them make better use of environmental sensors. A framework for the plan is expected to be ready for a ministers' conference on Earth observation in Tokyo next spring.



Satellite images, such as this of a Chilean mine, will form part of a new Earth-observation system.

NASA/ASTER

SAVE OUR CITY!

The Italian government is building a series of massive barriers to protect Venice from flooding. But scientists are still arguing over whether the plan will work, says Nicola Nosengo.

In the turbulent history of Venice, 4 November 1966 counts as one of the most notorious dates. Rain had been falling incessantly for two days, a prolonged low-pressure system had caused the sea level to rise dramatically, and a powerful wind forced wave after wave into the city's canals.

This exceptional coincidence of extreme meteorological conditions eliminated the tide's normal rhythm, and the water kept rising over three consecutive tides. By six o'clock that evening, it was almost two metres above average, and much of Venice was under water. Homes and businesses were devastated, and precious artworks destroyed. And it could so easily have been worse — had the wind blown for just a couple more hours, parts of the city would almost certainly have been swept away.

Venice sits in a lagoon, separated from the Adriatic Sea by a series of barrier islands. Originally, the lagoon consisted largely of mudflats, which helped to dissipate rising tides. But centuries of human intervention — including the diversion of rivers, widening of the lagoon's entrances and dredging of channels to accommodate shipping, and the draining of mudflats for construction and agriculture — have disrupted the lagoon's equilibrium with the sea. Today, the islands are eroded and the mudflats are largely submerged. When the sea surges, there is little to prevent the city's historic streets and squares from being inundated.

Although the events of November 1966



Water feature: advocates of the MOSE project hope that installing retractable barriers (above right) across the three inlets to Venice's lagoon (seen here from a satellite) will put an end to the city's flood misery (right).

were extreme, flood frequency had been increasing since the start of the twentieth century, and Venetians realized that something had to be done. Experts put forward various solutions, but arguments about their relative merits paralysed progress. In the meantime, people began to leave: since 1966, Venice's population has fallen from 127,000 to 65,000.

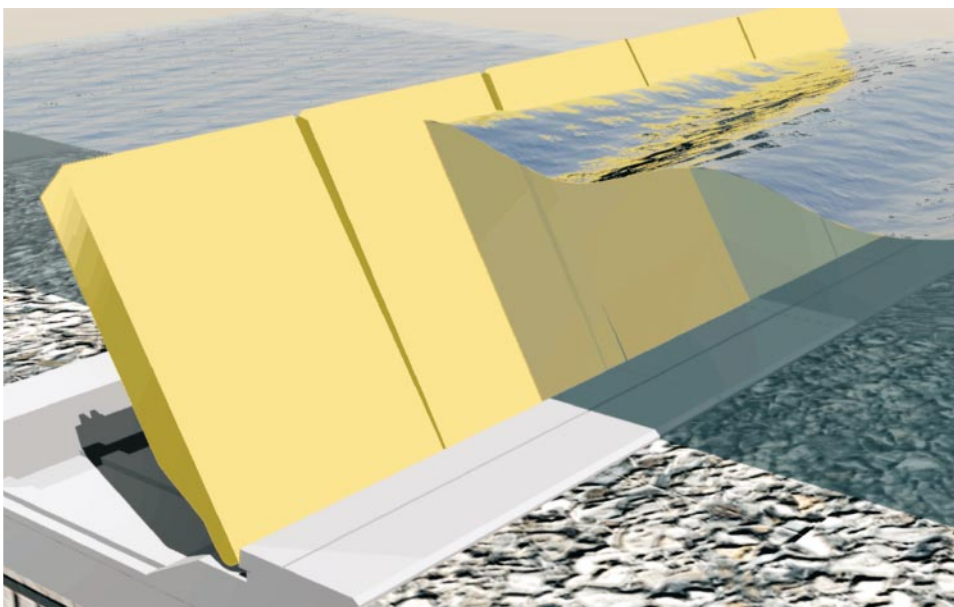
Costly solution

Finally, in December 2001, Italian prime minister Silvio Berlusconi declared that the most ambitious of the engineering solutions would put into action. Known by its Italian acronym of MOSE, the €2.3-billion (US\$2.6-billion) scheme will incorporate 78 hollow metal gates, each about 20 m high and 5 m thick, placed at the three main inlets of the lagoon. Each gate will lie flat on the sea floor under normal circumstances, with one end fixed to its foundations. In advance of particularly strong surges — exceeding 87 cm above the current average sea level — air will be pumped into the hollow gates, displacing the water inside them and causing the unfixed end to rise. Within an hour, the lagoon should be separated from the sea.

Construction of the gates is set to begin in 2006, and should be completed by 2011. In the interim, three stone reefs will be built outside each inlet to reduce the strength and speed of incoming tides.

The plan has been contested ever since it was proposed in the 1980s. Environmentalists claim that even brief interruptions of water exchange with the sea — MOSE's designers expect closures to last for four to five hours — could upset the lagoon's ecosystem. Many local politicians fear that the project will divert money needed for restoration within the city, and argue that 'softer' engineering solutions could be more effective. On 14 May this year, when work on the first stone reef began, Gianfranco Bettin, Venice's deputy mayor, reiterated his view of MOSE to the media: "Expensive, hazardous and probably useless."

Supporters and opponents of MOSE are still slugging it out almost daily in local newspapers and public debates, with scientists playing starring roles. One of the project's strongest critics is Venice-born Paolo Antonio Pirazzoli of the Laboratory for Physical Geography in Meudon, France, part of the CNRS, France's national research agency.



SIPA PRESS/REX FEATURES



The project has two main problems, Pirazzoli says. First, it was designed with an event similar to the 1966 flood in mind. But such a freak disaster is likely to occur no more than once every 165 years, Pirazzoli calculates. This makes MOSE a misdirected solution, he argues, as it will do nothing to combat the frequent minor floods that plague the city. "Last winter, there were more than 100 high-water events in Venice," says Pirazzoli. Many parts of the city were flooded before the water reached the 87-cm threshold, he says.

Rising discontent

What's more, claims Pirazzoli, MOSE will not be able to cope with the sea-level rise expected as a result of global warming. The Intergovernmental Panel on Climate Change predicts that seas will rise by an average of 50 cm over the next century, but MOSE is based on the assumption that the local sea level will rise by only 22 cm over the same period. Pirazzoli does not agree that the Adriatic Sea will be affected so mildly, and last year in *Eos*, the American Geophysical Union's newsletter, he argued that rising sea levels will render the barrier "obsolete"¹.

Pirazzoli believes the gates would have to remain raised for longer periods than their designers intended — up to several days at a time. Rain, drainage and the entry of seawater through gaps between the gates would then cause the lagoon's water to rise. In January, Pirazzoli and Georg Umgiesser, an oceanographer at the Institute for the Study of Large-Mass Dynamics in Venice, part of the CNR, Italy's national research council, published an analysis of five historical surges, including the 1966 flood. They added 50 cm to their model's sea level and concluded that, in each case, the gates would not have prevented flooding, because of water accumulating in the lagoon².

Pirazzoli favours softer interventions to control regular floods. These include reducing the depths of the inlets by raising their beds, and narrowing the large canals that enter the city's harbour. The idea is to impede the flow of rising tides. If these measures fail, Pirazzoli argues that it may be necessary to separate the lagoon completely from the sea, apart from a few narrow canals — much as was done between 1924 and 1932 in the Netherlands, when a dike was erected along the country's coast, blocking off the North Sea and creating a huge freshwater lake called the IJsselmeer.

MOSE's supporters contest Pirazzoli's assumptions. "His analysis is completely flawed," claims Andrea Rinaldo, a hydraulic engineer at Padova University, near Venice, who was a member of an international expert group asked to assess the project for the Italian government³. Like Pirazzoli, Rinaldo was born and grew up in Venice, and the 1966 flood left a deep impression on him. "Furniture was floating around my flooded room," he recalls.

Rinaldo argues that water will not rise as rapidly as Pirazzoli assumes when the gates are closed. MOSE will not prevent occasional floods in the city's lowest parts, Rinaldo concedes, but he argues that without intervention,

the 1966 calamity could be repeated at any time. "We should not be too confident of living in a world in which the probability of extreme events is minimal," he says. "Not protecting the city would be irresponsible."

Rinaldo, who with other members of the international expert group published a response to Pirazzoli's criticisms in the same issue of *Eos*⁴, argues that his opponent overestimates the extent to which today's flooding is a consequence of the construction of shipping channels. Rinaldo points instead to the fact that Venice has subsided by 13 cm as a result of pumping water from the aquifers beneath the city from the 1950s to the 1970s.

Species barrier

Even if MOSE provides effective flood defence, the question remains of how it will affect the lagoon's environment. During construction, five million cubic metres of sediment will be excavated from the areas next to the inlets, which are the richest in species diversity. Environmentalists fear that the lagoon's sea grasses, on which its ecosystem is based, will be choked by mud. They also argue that prolonged gate closures will exacerbate problems with the build-up of sewage.

A team at the National Institute of Oceanography and Experimental Geophysics in Trieste is now assessing the risk of increased pollution. Their radar measurements of sea currents suggest that the lagoon will rapidly be cleansed by the tide even after gate closures lasting up to five days⁵. "What became quite clear is that water is exchanged between the sea and the lagoon at an extremely fast rate," says Miroslav Gacic, who led the study. "When the tide enters the lagoon, it brings in new, oxygenated water. It is not the same water that came out six hours before." Indeed, Gacic argues that Pirazzoli's preferred solution of restricting the lagoon's inlets entails greater risk, as it would alter the exchange of water between the lagoon and the sea.

But MOSE's critics remain defiant in what has become a politicized debate: advocating MOSE implies support for Berlusconi's right-wing government; criticism is seen as backing for the left.

Scientists hope that an international meeting on Venice's flooding problems, planned for September in Cambridge, UK, will restore some objectivity. But as the arguments rage, it seems that the protagonists can agree on only one thing: their hope that the nightmare of 4 November 1966 doesn't return to haunt Venice before some means of protecting this historic treasure has been put in place. ■

Nicola Nosengo is an intern with *Nature* in Munich.

1. Pirazzoli, P. A. *Eos* **83**, 217, 223–224 (2002).
2. Pirazzoli, P. A. & Umgiesser, G. *CNR Technical Report* **256** (2003).
3. Collegio di Esperti di Livello Internazionale. *Report on the Mobile Gates Project for the Tidal Regulation at the Venice Lagoon Inlets* (1998).
4. Bras, R. L. *et al.* *Eos* **83**, 217, 224 (2002).
5. Gacic, M. *et al.* *Eos* **83**, 217, 221–222 (2002).

♦ <http://ccru.geog.cam.ac.uk/events/venice2003>

Vital statistics

DNA microarrays have given geneticists and molecular biologists access to more data than ever before. But do these researchers have the statistical know-how to cope? Claire Tilstone investigates.

In June, Nick Fisher issued a dire-sounding warning. He is president of the Statistical Society of Australia, and fears that sloppy statistics could undermine the revolution promised by genomics and biotechnology. "If the collection, analysis and interpretation of the data are flawed then it may not only be a waste of a valuable resource — we could draw faulty conclusions and potentially risk our health and environment," Fisher claimed in a release to the media.

Fisher has a vested interest: he wants research organizations to employ professionally accredited statisticians — his society's members — to oversee the collection and analysis of genomic data. But he has a point. Technologies such as DNA microarrays have sent avalanches of data tumbling into labs where, previously, analysing the results of an experiment meant little more than glancing at an electrophoretic gel.

Over the past few years, microarrays, also known as DNA chips, have transformed molecular genetics, allowing researchers to study the activity of thousands of genes at a time. For example, you can compare healthy tissues with those that are cancerous, to identify genes that become more, or less, active in a developing tumour. But how do you separate significant differences in gene expression from background fluctuation?

This is where good experimental design and statistical analysis should come in. But there are no simple answers: interpreting microarray experiments is taxing the skills of even the most adept number-crunchers. "It's a technical and esoteric topic," says Paul Meltzer, a cancer geneticist and microarray specialist at the National Human Genome Research Institute in Bethesda, Maryland. If trying to make sense of microarray data has left you with spots circling before your eyes, you're in good company.

The problem is perverse: a typical microarray experiment provides both too much information, and too little. In most research projects, the idea is to study a small number of variables and repeat the measurements over and over again. Provided that you perform enough replicates, standard statistical tests can establish whether experimental results have real significance, or are more likely to be a consequence of random noise. Microarrays turn this approach on its head: there can be thousands of variables, corresponding to the number of individual genes

being studied; but the high cost of the chips means that the number of repeated observations is usually very low.

In their initial forays into microarray research, many biologists didn't even try to use statistical methods. Some of the earliest papers simply recorded whether genes were active or not. Even when researchers began to use microarrays to measure levels of gene expression, they tended not to quote standard error values or confidence limits, and differences were judged to be meaningful if they exceeded some arbitrary level. Today, many manuscripts submitted to journals still focus their conclusions on genes that show a change in activity of, say, more than twofold.

But how do you know whether or not an apparent twofold change in gene expression is biologically meaningful? That's a difficult question to answer, because of the many

sources of noise that can cloud the results.

The concept of microarray analysis is simple enough: messenger RNAs are extracted from a biological sample, converted into DNA, labelled with fluorescent dyes, and then washed over a glass slide bearing a grid spotted with DNA sequences from known genes. The labelled sequences bind to spots representing the genes from which the messenger RNAs were transcribed. So by analysing the location and intensity of the fluorescent signals, you can determine the level of activity for each gene. In some cases, it is possible to analyse two samples on the same chip by using different coloured dyes.

But noise creeps into microarray experiments at every stage, from the preparation of tissue samples to the extraction of data. Using different dyes can influence the results recorded by the lasers that measure the fluo-



The main deterrent is cost: commercial DNA chips retail for about US\$1,000 each. And in some cases, for instance when studying rare diseases, it can be difficult to obtain enough samples to perform the desired replicates. One way round the cost issue is to collect several samples, pool them, and apply them to just one microarray. But the jury on pooling is still out. Many experts believe it is a bad idea because valuable information on sample-to-sample variation is obscured;



A sight for sore eyes? Statistical analysis of the data provided by DNA chips can be a major headache.

Even if you decide to perform proper, non-pooled replicates, it is difficult to know how many to do. One paper investigating the topic, published in 2000, recommended at least three replicates². But among statisticians who have considered the issue, there is no clear consensus. “The number of replicates really depends on the kind of accuracy one wants to achieve,” says Ernst Wit, a statistical geneticist at the University of Glasgow.

Biologists are now trying to use replication and statistical analysis to separate meaningful changes in gene expression from background noise, often using software packages produced by academic researchers and available for free download, or marketed by chip manufacturers. But in many cases, say experts, the statistics aren't being used correctly. "The majority of microarray papers are analysed with substandard methods," claims David Allison, a biostatistician at the University of Alabama at Birmingham.

One source of confusion is how to correct for the 'false positive' results that are a consequence of the multiple comparisons inherent to microarray analysis. The results for an individual gene may suggest that there is only a 5% chance that recorded differences in its activity are the result of chance fluctuations; but if you repeat the same test across 10,000 genes, you are likely to get 500 'significant' results even if there is no real difference in gene expression. For simpler data sets, established methods exist for dealing with multiple comparisons. But opinions differ on how these should be applied to microarray data, says John Quackenbush, a specialist in genomic data analysis at The Institute for Genomic Research in Rockville, Maryland.

Grouping genes in this way involves statistical techniques collectively known as cluster analysis. The first step is to draw up a gene-expression matrix in which the rows represent individual genes, the columns are individual samples, and each cell contains a measure of



the gene's activity. From this matrix, each gene can be given a coordinate called its 'expression vector', which represents a point in an n -dimensional mathematical space, where n is the number of samples. The precise position of each gene in each dimension depends on its level of activity in the sample concerned.

The genes are then clustered into groups by methods that 'measure' the distances between their respective expression vectors. The difficulty is that there are various ways of doing this, which delight in exotic names such as 'k-means clustering' or 'self-organizing maps'³. Statistical experts struggle to agree over which clustering method should be applied under what circumstances. And for most biologists — even those whose eyes haven't glazed over at the initial mention of an n -dimensional space — the detailed workings of these methods remain a mystery.

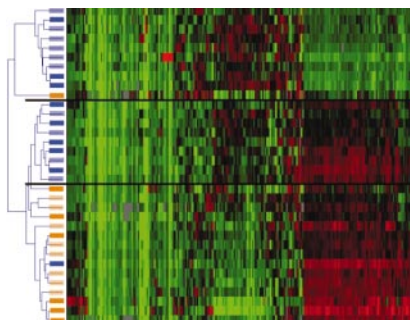
Nevertheless, cluster analysis has caught on, and the most common approach is a technique called hierarchical clustering, first applied to microarray data in a paper published in 1998 by David Botstein and other microarray pioneers at Stanford University in California⁴. Here, an iterative algorithm draws up a tree in which the lengths of the branches correspond to the degree of similarity between genes.

General clusters

Hierarchical clustering has helped flustered researchers to make sense of what would otherwise be unmanageable spreadsheets of data. But statistical purists point to problems with the way in which it is being applied. Some argue that the technique is best suited to determining relationships between a small number of variables, rather than deriving patterns involving thousands of genes across a huge data set. "Hierarchical trees are famously unreliable for good high-level clusters," explains Wit.

This means that biologists can be sent down blind alleys, if they see clusters that reinforce their own assumptions about the relationships between individual genes. "People just tend to pick their favourite cluster for further study," says Allison, who likes to play a trick on biologists: he presents them with two trees drawn up from simulated microarray data, one representing genuine clusters of genes with similar expression profiles, the other in which the genes have been clustered randomly. When he shows this slide, Allison is greeted by gasps and chuckles, as those in the audience realize that they don't have an intuitive ability to recognize a 'true' cluster.

Cluster analysis can also group samples that show similar patterns of genome-wide gene expression. The power of this approach was demonstrated in 2000 by a team led by Louis Staudt of the National Cancer Institute in Bethesda, Maryland, which used hierarchical clustering to group B-cell lymphomas into two distinct classes. These correlated with dif-



Branching out: cluster analysis can group samples that show similar patterns of gene expression.

ferences in patient survival, and seemed from their gene-expression profiles to be related to the stage of development of the cell from which the cancer originated⁵. But last year another group, using a different clustering method, failed to find this association⁶.

Experts argue that such disagreements are only to be expected, given that microarray analysis is a young and fast-moving field. They point to efforts to refine the statistical methods applied to microarray data, such as the annual Critical Assessment of Microarray Data Analysis (CAMDA) meeting, now in its fourth year. Organized by Simon Lin and Kimberly Johnson of Duke University in Durham, North Carolina, CAMDA culminates in the award of a prize to the group judged to have conducted the best analysis of real data sets posted before the meeting on CAMDA's website.

Quackenbush, who is giving a seminar at the next CAMDA gathering in November, suspects that there may never be agreement on the 'right' statistical techniques to deploy on microarray data. The future, he suggests, lies in incorporating additional biological information into the analysis. He points to a paper published in March, in which microarray data and information on the chromosomal location of genes that influence obesity were considered together to identify two subtypes of obesity in mice⁷.

Array of hope

While microarray experts push back the frontiers, efforts are being made to improve standards of experimental design, data presentation and analysis among rank-and-file users of the technology. Particularly useful is a set of guidelines called minimum information about a microarray experiment (MIAME), laid down by the international Microarray Gene Expression Data Society. These include specific statements about experimental design, including the number of replicates done, allowing researchers to interpret one another's data more easily.

Many journals are now toughening up their criteria for accepting papers describing microarray experiments. Since December 2002, for instance, *Nature* and its sister research journals have required authors of

such papers to complete a MIAME checklist; data must also be submitted before publication to one of the two main public repositories for microarray data⁸. One of the most explicit statements has come from the journal *Arthritis & Rheumatism*, which in April last year published guidelines stressing the importance of appropriate statistical analysis⁹. "It is not sufficient to say that the expression of a particular gene is twofold greater in a sample than in a control," the journal's editors stated.

Chip manufacturers and specialists in academia are also trying to help microarray users become more rigorous. Last year, for instance, the chip-making firm Affymetrix of Santa Clara, California, began a series of worldwide workshops on experimental design and the use of statistical software. Meanwhile, the Bioconductor project, run by the Dana Farber Cancer Institute in Boston, is backing its programs for genomic data analysis with short courses in their use. By the end of 2003, these will have taken place in the United States, Europe and Taiwan.

Standards should improve further as statistics and bioinformatics become more prominent components in the training of molecular biologists, and as a growing number of statistical experts gets to grips with the complexities of microarrays. "There are lots of good statisticians out there, but not as many of them have been exposed to microarray data as are needed," says Meltzer.

In the meantime, the message to biologists is clear: if you want to work with microarrays, you need to find yourself one of these precious experts — and don't wait until after you've collected your data. The following advice, from pioneering British geneticist and statistician Ronald Fisher, rings even more true today than when he uttered it, back in 1938: "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

Claire Tillstone is a postgraduate at the University of Bath, UK, studying science communication; further reporting by Peter Aldhous, *Nature's* chief news and features editor.

1. Kuo, W. P., Jensen, T.-K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. *Bioinformatics* **18**, 405–412 (2002).
2. Lee, M.-L. T., Kou, F. C., Whitmore, G. A. & Sklar, J. *Proc. Natl Acad. Sci. USA* **97**, 9834–9839 (2000).
3. Quackenbush, J. *Nature Rev. Genet.* **2**, 418–427 (2001).
4. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
5. Alizadeh, A. A. et al. *Nature* **403**, 503–511 (2000).
6. Shipp, M. A. et al. *Nature Med.* **8**, 68–74 (2002).
7. Schadt, E. E. et al. *Nature* **422**, 297–302 (2003).
8. *Nature* **419**, 323 (2002).
9. Firestein, G. S. & Pisetsky, D. S. *Arthritis Rheum.* **46**, 859–861 (2002).

Critical Assessment of Microarray Data Analysis

♦ www.camda.duke.edu

Microarray Gene Expression Data Society

♦ www.mged.org

Affymetrix

♦ www.affymetrix.com

Bioconductor

♦ www.bioconductor.org

Flawed science underlies laws on transgenic crops

United States needs a firm basis for commercialization, not a trade war with Europe.

Sir — Your News story “Trade war looms as US launches challenge over transgenic crops” describes the conflict between the United States and the European Union and reports questions over the labelling and traceability of transgenic crop products (*Nature* **423**, 369; 2003). I would like to add that commercialization and cultivation of transgenic seeds are also in dispute.

The US government has often stated that its consent to large-scale commercialization was based on sound science, as required by the United Nations’ Cartagena Protocol on Biosafety and other international agreements. However, examination of documents of the US agencies responsible, as well as confidential material, reveals a

poor scientific basis for this decision^{1–3}.

Two US National Research Council publications^{1,2} document the superficial and flawed nature of many scientific conclusions, revealing the lack of long-term risk assessment and of post-release monitoring. Superficial assumptions, poor experimental design and statistical flaws are also described in a report on insect-resistant (*Bt*) plants³. The conclusion of a broad-scale analysis⁴ is that key experiments on both the environmental risks and benefits of genetically engineered plants are lacking.

In view of these deficiencies, a transgenic trade war seems ill-founded. Instead, the priority should be the generation of a sound scientific basis for

registration of transgenic crop plants. Processes to be clarified at the laboratory and field levels include recombination and gene transfer events, as well as environmental selection processes that could cause resistance development and population shifts, with many potential secondary consequences.

H. Sandermann Jr

*Institute for Biochemical Plant Pathology,
GSF-National Research Centre for Environment
and Health, D-85764 Neuherberg, Germany*

1. National Research Council *Genetically Modified Pest-Protected Plants: Science and Regulation* (National Academies Press, Washington DC, 2000).
2. National Research Council *Environmental Effects of Transgenic Plants: The Scope and Adequacy of Regulation* (National Academies Press, Washington DC, 2002).
3. Marvier, M. *Am. Sci.* **89**, 160–167 (2001).
4. Wolfenbarger, L. L. & Phifer, P. R. *Science* **290**, 2088–2093 (2000).

Getting to the heart of transpiration in plants

Sir — Melvin T. Tyree in his Concepts essay on plant hydraulics “The ascent of water” (*Nature* **423**, 923; 2003) mentions the well-known comparison of the mammalian heart, the pump circulating the blood, to the movement of water in plants by “the pulling force generated at the evaporative surface of leaves” (transpiration). One interesting question arising from this comparison is whether transpiration is as essential for a plant as is the heart for an animal?

Although it is generally accepted that transpiration for plants is essential, doubts have often been raised. For example, Paul Kramer on page 293 of his book *Water Relations of Plants* (Academic, Orlando, 1983) states: “transpiration can be best regarded as an unavoidable evil” — unavoidable because no plant ‘skin’ has evolved that is permeable to CO₂ but impermeable to water vapour. Which view is correct?

There are several distinct aspects to long-distance water transport. First, water lost via the evaporative surface of plant leaves has to be replenished to prevent the plant wilting, as outlined by Tyree. Second, ‘growth water’ has to be lifted to the top of plants (including trees) for the expansion growth of leaves, fruits and shoots. And third, all the water flowing downwards in the phloem has to be replaced by an equal volume of water moving upwards.

Only the water flow in the first of these is transpiration-driven. It does not take place if transpiration is zero or almost zero (if relative humidity is 100% or at night, when the stomata are closed). Evaporation generates a pulling force sufficient to replenish the amount of water lost by

transpiration, thus minimizing the problem it creates for the plant.

The water movements caused by the second two processes do not depend on transpiration. The evaporative loss of, say, 100 molecules of water could at the very best generate a force sufficient to lift 100 molecules, but not 110 (those lost by evaporation plus those required for growth). So how is the amount of water in the second two processes moved against gravitational force? By the difference in water potential created by growth, volume flow of assimilates and root pressure. H. Beevers and I have estimated that in less than four days a negative water potential large enough to pull water up a 100-m tree would arise in the absence of any transpiration (W. Tanner and H. Beevers *Proc. Natl Acad. Sci. USA* **98**, 9443–9447; 2001). Transpiration is not essential for net water movement: that is, the amount of water the plant requires. This calls into question the analogy with the mammalian heart.

Widmar Tanner

*Lehrstuhl für Zellbiologie und Pflanzenphysiologie,
Universität Regensburg, 93040 Regensburg, Germany*

Natural decaf could brew trouble for farmers

Sir — The reduced-caffeine transgenic coffee plant (S. Ogita *et al.* *Nature* **423**, 823; 2003) must be one of the first transgenic plants where a natural plant compound is suppressed. Caffeine has been proposed to form part of the plant’s defence mechanisms, specifically against insects (J. A. Nathanson *Science* **226**, 184–187; 1984) and slugs (R. Hollingsworth, J. Armstrong

and E. Campbell *Nature* **417**, 915–916; 2002). These transgenic plants may thus be more successful in testing this hypothesis than as an agricultural crop.

It is interesting that the transformation was carried out on *Coffea canephora*, the lower-grade robusta coffee, which has about twice as much caffeine as arabica has. As its name suggests, robusta is more resistant to attack by insects and diseases and it is likely that caffeine plays a role in that.

The coffee industry may be especially keen to reduce caffeine in robusta, as over the years it has gradually been slipping more of it into standard arabica blends. In the opinion of Ernesto Illy, chairman of the Institute for Scientific Information on Coffee and espresso *maestro di tutti i maestri*, this has contributed to the present sluggish demand for coffee world-wide as consumers drink less to maintain a constant caffeine intake (*Coffee and Cocoa International* **28**, no. 6, 20–22; 2001).

Because coffee is a perennial plant that takes about three years to come into production and may stay in the ground for 20 or more years, changing to a new variety is a major investment. Farmers would need assurance that the new plant is as resistant to attack as other varieties. Unfortunately, it could take many years for an agent to adapt to the new plant, so short-term tests might be insufficient. Further, coffee is grown in many regions and habitats, so very widespread field trials would be needed.

Hence, although a low-caffeine plant would be a useful research tool, there could be unintended long-term consequences most acutely felt by the poor farmer.

P. S. Baker

*Coffee Projects Coordinator, CABI Commodities,
Bakeham Lane, Egham TW20 9TY, UK*

Beyond the skin-bag

The brain extends its reach outside the body, so are we all cyborgs?

Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence

by Andy Clark

Oxford University Press: 2003. 229 pp. \$26

Don Ihde

Andy Clark illustrates his point that human intelligence is distributed globally far beyond the limits of the “ancient biological skin-bag” by referring to online bookseller amazon.com’s practice of noting that people who bought books by this author also bought books by that one. So was it by coincidence, I wonder, that the very week that I was asked to review *Natural-Born Cyborgs*, Clark’s name popped up under amazon.com’s listing of my own book *Bodies in Technology*? Coincidence or not, I found *Natural-Born Cyborgs* to be simultaneously highly interesting, provocative and easy to read.

The term ‘cyborg’ is a portmanteau of ‘cybernetic organism’, coined by Manfred Clynes and Nathan Kline in 1960 and popularized by Donna Haraway in her essay *Cyborg Manifesto*. Here it is used to address the issue of human intelligence and embodiment, particularly with respect to the technologies that increasingly hybridize humans and machines. Clark, who heads the cognitive-science programme at Indiana University, is a second-generation cognitivist who has gone beyond the early logicism of artificial intelligence. He explores the notion that humans and their technologies have always been so closely interrelated that we have always been cyborgs. Now, however, by including technologies that are worn on the surface of our bodies or even implanted within them, we are beginning to open new dimensions for human intelligence. Clark’s own perspective on these developments is dominantly optimistic, although he does include a list of worries in a chapter entitled “Bad borgs?”

I find Clark’s position fascinating philosophically because he sometimes retains the cartesian ‘consciousness-in-a-box’ notion, which is now nearly 400 years old. Recently reinvented as a ‘brain-in-a-vat’ by Daniel Dennett, whom Clark often follows, this idea implies the existence of a homunculus — an all-controlling self — that is now simply relocated into a brain. Much cognitive science and some neurology support this highly localized notion of consciousness, the origins of which lie in René Descartes’ use of the camera obscura as the model for the body and mind. But Clark also uses what I would call a phenomenological set of descriptions,



Many hands make light work for performance artist Stelarc, as he demonstrates his third arm.

in which our bodily self-experience goes well beyond the ‘biological skin-bag’ — a theme that Clark seems to get from the philosophers Hubert Dreyfus and Maurice Merleau-Ponty. Clark imaginatively, but sometimes inconsistently, combines these two drastically different ideas.

Clark recognizes that our bodily experience extends beyond the limits of the ‘skin-bag’ by means of tactility at a distance, through, as Merleau-Ponty suggested, the blind man’s cane and the sense of bodily extension felt when driving a car. But the newer interfaces, both implants and wearable devices (particularly electronic ones), vastly extend these possibilities. Historically, many technologies were first designed as prosthetic devices, such as telephones for helping those with hearing difficulties, or typewriters for the visually impaired. New prostheses include artificial vision devices such as the ‘Dobelle eye’ and the ‘tactile visual sensory substitution’ system.

The book discusses both implant technologies, which are often still primitive, and experience-enhancing ‘wearable’ technologies. For example, the performance artist Stelarc wears a third hand that he can use for writing and gesturing. The array of such technologies discussed in *Natural-Born Cyborgs* is impressive and entertaining, giving the book a potentially wide audience that includes those interested in cognitive science, performance art and the philosophy of mind.

Clark’s wide-ranging interests and knowl-

edge of cutting-edge technologies also help him to overcome some of the traps (and hype) into which older artificial-intelligence, virtual-reality and techno-utopian fantasies often fall. Technologies as diverse as prosthetic cochlear implants and the Internet (discussed extensively in the chapters entitled “Where are we?” and “Global swarming”) are shown as enhancements and extensions of human experience and intelligence. This is not to say that all implants provide such utilitarian enhancement. For example, Clark mentions a medical implant that triggers orgasms, and the information society, which brings many benefits, also causes problems through its increasing ability to keep tabs on us.

When reading *Natural-Born Cyborgs* there were moments when my tendency to criticize was piqued — for example, when Clark maintains a homunculus version of conscious brain activity, complete with a desire to control things centrally by giving commands, thus falling again into cartesian antiquity. Yet, by then recognizing that learning raises such technologies to skill levels not hindered by discrete commands, and by his emphasis on plasticity, Clark leads us to a sense of embodiment beyond the antique.

In the end, I was quite happy with the amazon.com dissolution of my degrees of separation from Andy Clark.

Don Ihde is in the Department of Philosophy, Stony Brook University, Stony Brook, New York 11794-3750, USA.

Tea, cake and computers

A Computer Called LEO: Lyons Teashops and the World's First Office Computer

by Georgina Ferry

Fourth Estate: 2003. 240 pp. £15.99

Anthony Ralston

At first glance it seems astonishing that a company known mainly for operating cafes that served tea and cakes in Britain should not only have installed the first ever computer for business applications, but should also have built it. But as Georgina Ferry makes clear in this book, it wasn't really so surprising because J. Lyons & Co. had become, under the leadership of John Simmons, a pioneer in the application of technology to management in the years before the Second World War.

It was therefore natural that two senior Lyons employees, Raymond Thompson and Oliver Standingford, should have been sent to the United States in 1947 to study recent developments in office automation and to learn what they could about ENIAC (the Electronic Numeric Integrator and Computer), which had just been completed at the University of Pennsylvania. (ENIAC was not the first electronic computer — the Colossus of Britain's Bletchley Park deserves that accolade. However, Colossus was designed for a special purpose, codebreaking, and was still a secret in 1947. Indeed, even in the late 1970s, I could not find a British author to write an article on Colossus for an encyclopaedia that I was editing because it was still covered by the Official Secrets Act.)

Thompson and Standingford returned to Britain imbued with the potential of computers for business management. They had learned that, in the United States, only the Prudential Insurance Company was thinking along similar lines. In late 1947 the pair visited the University of Cambridge, where Maurice Wilkes was building the EDSAC (the Electronic Delay Storage Automatic Calculator), and they knew they had found what they needed. Although EDSAC had been designed with scientific and engineering calculations in mind, they believed that its design could be adapted to Lyons' needs.

How this could be done wasn't at all obvious. Most scientific applications tended to be 'compute-bound' — lots of computing and little input or output — whereas business or data-processing applications tended to be 'input/output bound', with lots of input and output but very little computing. Could a computer intended for the former be adapted for the latter? The Lyons management convinced themselves that it could, although serious problems with input and output



Mane frame? LEO, the first business computer, was set up and run by J. Lyons & Co., which operated a chain of tea shops in Britain.

bedevilled the LEO (Lyons Electronic Office) project for much of its lifetime.

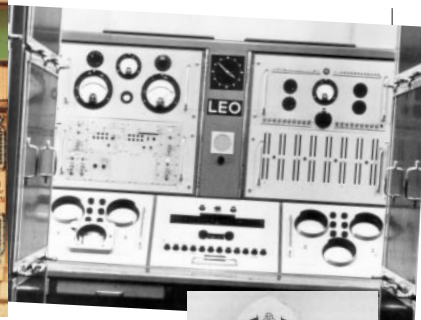
LEO was, essentially, a copy of the EDSAC with a few hardware innovations and some limited software advances. Its main claim to fame is in the applications area. On 29 November 1951, LEO ran the first ever business application, a program to calculate the value of Lyons' output of bread, cakes and pies. This was long before any similar application in the United States, Prudential's use of computers having been delayed by hardware problems. Over the following years, LEO took over much of Lyons' data-processing work, and was used for some non-Lyons applications, including various defence-related calculations. When LEO executed defence programs, the Official Secrets Act required it to be encircled by red tape, with only authorized personnel allowed near the computer.

It was particularly quixotic of Lyons not only to build and install their own machine but also to go into the business of building and selling computers to others. An updated version, LEO II, still used valves; LEO III was redesigned to incorporate semiconductor technology. Ten LEO IIs were delivered to customers between 1957 and 1961, and 61 LEO IIIs were sold, mainly from 1962 to 1966. By then, the manufacturing and marketing power of US companies, particularly IBM, was severely impairing the ability of Lyons to compete as an independ-

ent manufacturer of computers. In 1963, Leo Computers was merged with — but effectively taken over by — the computer division of English Electric, in the first of a series of mergers that eventually led to the creation of the last major British computer company, International Computers Ltd.

The story of LEO has already been told in two more technical books: *LEO: The First Business Computer* by Peter Bird (Hasler, 1994) and *LEO: The Incredible Story of the World's First Business Computer* by David Caminer *et al.* (McGraw-Hill, 1997); further information can be found on the web at www.leo-computers.org.uk. But Georgina Ferry tells the story well, with less technical detail but with nice capsule histories of the development of computing in Britain and the United States, particularly in the decade after the end of the Second World War. Her book is rather more accessible to the general reader than the others, and is a good read for anyone interested in LEO or, more generally, in postwar computing in Britain. ■

Anthony Ralston is professor emeritus of computer science and mathematics at the State University of New York at Buffalo, and an academic visitor in the Department of Computing, Imperial College, London SW7 2BZ, UK.



The struggle for sexual inequality

Sex Wars: Genes, Bacteria, and Biased Sex Ratios

by Michael E. N. Majerus

Princeton University Press: 2003. 280 pp.

\$45, £29.95

Göran Arnqvist

In the first edition of *On the Origin of Species*, Charles Darwin made a brave and essentially correct attempt to explain the striking rule that males and females occur

in equal numbers. In later editions, however, he described the problem of the even sex ratio as being "so intricate that it is safer to leave its solution to the future". Ronald Fisher took up this challenge in 1930, noting that every offspring has one mother and one father. Therefore, the sexes on average fare equally well in passing their genes on to the next generation under an even sex ratio. Fisher realized that a mother that produces an excess of the rarer sex (whichever this may be) will be favoured by natural selection. So, like a pendulum in motion that eventually comes to rest, the sex ratio will stabilize at equal numbers of the sexes over evolutionary time. But there are many

exceptions to this rule, and understanding them has proved to be a challenge for evolutionary biologists.

These exceptions form the centre of gravity of *Sex Wars*. Michael Majerus of the University of Cambridge starts off by walking us through the background and some of the basics: asexual versus sexual reproduction, sexual selection, the sex-determining mechanisms, the sex ratio and the various 'economic' factors that can lead to uneven sex ratios are all covered in the opening sections. But the bulk of the book is about the various genetic conflicts that knock the sex ratio off its evolutionary equilibrium.

For a gene located on a nuclear autosome — any chromosome except the sex chromosomes — life is as good in a male as it is in a female, in terms of the probability of being transmitted to the next generation. For extranuclear genes, however, males are a genetic dead-end because the sperm contributes nothing but its nuclear genes to the fertilized zygote. Cytoplasmic genes located in mitochondria or in endocellular microorganisms are expected to favour the female line, through which they are transmitted to the next generation.

This sets the stage for a battle between autosomal genes on one hand, and genes with unequal transmission rates in the two sexes on the other. I found the portrayal of these opponents by Majerus very exciting, and his account of them offers many insights. These evolutionary law-breakers use a variety of intrusive tactics (including killing sperm containing Y chromosomes, transforming genetic males into functional females, or simply killing males), which all bias the sex ratio towards females, sometimes even to the point when the 'host' is driven close to the brink of extinction because of a lack of males in the population. The autosomal genes fight back, however. Majerus describes several cases of autosomal 'rescue' genes that counter the effects of the law-breakers. The natural history of sex-ratio distortion is the definite strong point of this book — Majerus tells many intriguing and entertaining tales about various reproductive curiosities.

No book can be written for everyone, but every book should be written for someone. This book is aimed at "a wide audience". Any reader of this book will find much of interest, but will also be frustrated. Like the indecisive Buridan's ass who could not choose between two equal bales of hay and therefore starved to death, Majerus roams the treacherous no-man's-land that lies between the general reader and the specialist, unable to quite reach either. General readers will stumble over many of the terms and turns in the more detailed and complicated sections. There is a glossary at the end of the book, but it is incomplete and partly confused.

Specialist readers, on the other hand, will



Sex to die for: females of the ladybird *Adalia 10-punctata* distort the sex ratio by killing males.

find the lack of reference to many case studies and ideas, as well as some more technical problems, quite annoying at times. They might also find that the book loses the battle with a more exhaustive book on sex ratios — *Sex Ratios*, edited by I. C. W. Hardy (Cambridge University Press, 2002) — that recently reached the bookshelves. If *Sex Wars* were less technical and more 'popular', it might have found more enthusiastic readers. ■

Göran Arnqvist is in the Department of Animal Ecology, Evolutionary Biology Centre, University of Uppsala, Norbyvägen 18d, SE-752 36 Uppsala, Sweden.

..... The next happy pill

Better Than Prozac: Creating the New Generation of Psychiatric Drugs

by Samuel Barondes
Oxford University Press: 2003. 240 pp.
\$26, £16.99

Les Iversen

The past 50 years have seen a transformation in the practice of psychiatry. The emphasis has moved from psychotherapy with a limited repertoire of drugs that were used only in the most severely ill patients, to the widespread use of drugs that act on the central nervous system (CNS), with psychotherapy being an expensive luxury for a minority of patients. Samuel Barondes is well placed to review the psychopharmacology revolution and to offer predictions for the future, as he has worked through this period as a practising psychiatrist and as an active researcher who has made important contributions to the field.

An early section of the book explains how the existing drugs used to treat psychiatric illnesses — antidepressants, antipsychotic drugs, tranquillizers and amphetamines — were discovered, with colourful accounts of the individuals involved. Along the way there

New in paperback

Mapping Human History: Genes, Race, and Our Common Origins

by Steve Olson
Mariner, \$14; Bloomsbury, £7.99

Our Posthuman Future: Consequences of the Biotechnology Revolution

by Francis Fukuyama
Profile, £8.99

The Golden Ratio: The Story of Phi, The Extraordinary Number of Nature, Art and Beauty

by Mario Livio
Review (Headline), £7.99

Nexus: Small Worlds and the Groundbreaking Theory of Networks

by Mark Buchanan
W. W. Norton, \$14.95
"Mark Buchanan gives a cogent and engaging description of recent developments in complex networks." Sidney Redner *Nature* **418**, 127–128 (2002).

The Monkey in the Mirror: Essays on the Science of What Makes Us Human

by Ian Tattersall
Harvest Books, £13

are explanations in simple language of how the drugs affect various neurotransmitter mechanisms in the brain, and an introduction to the concepts of controlled clinical trials, the placebo effect and the mystery of why it takes many weeks of treatment before the maximum therapeutic benefits of many CNS drugs are seen.

Barondes uses individual case histories to illustrate many of these points, a device that helps to involve the reader with real people who have benefited from drug treatments, and to illustrate some of the shortcomings of the medicines that are presently available. There is a graphic description, for example, of the power of the selective serotonin re-uptake inhibitor (SSRI) Prozac to transform a patient from a state of sadness and agitation to one of calm and tranquility, albeit with a diminished intellectual passion and reduced sex drive.

A second section provides an overview of the prospect that a better Prozac will emerge from current research in this field. Like many others, the author believes that this will come from a better understanding of the nature of the biological disturbances that underlie psychiatric illnesses. There is reason for optimism: several valuable clues have already emerged from research on the genetic risk factors underlying these illnesses, and potential candidate genes have been

described in the past year for schizophrenia and for manic-depressive illness. It is likely that most psychiatric illnesses are partly inherited through genetic risk factors but also depend on environmental factors. Nevertheless, identifying the genetic risk factors may provide important clues about new treatment strategies.

A case in point is Alzheimer's disease, in which mutations in several different genes are known to increase the risk of developing the disease, and in a small number of patients, mutations in the amyloid precursor protein (APP) cause the illness directly. This new genetic information has greatly strengthened the hypothesis that the abnormal formation and deposition of amyloid- β , a protein derived from APP, is a key event underlying the pathology of Alzheimer's. This in turn has triggered a large research effort directed towards the discovery of drugs that can reduce the formation or deposition of amyloid- β . Whether these will prove effective in the clinic remains to be seen, but for the first time a rational approach to the treatment of this devastating illness has become possible.

Understanding the underlying disease process and developing medicines that interrupt this process remain the ultimate goals in devising a better Prozac. In the meantime, better drugs can be developed to treat the symptoms of psychiatric illnesses, for example by the introduction of antagonists of the neuropeptide substance P to create new antidepressants that lack the adverse effects of SSRIs on sexual function. Alternatively, antagonists of the neuropeptide corticotropin-releasing factor could be used to make potential 'anti-stress' agents. New uses are also being found for existing drugs — for example, SSRIs can be used in the treatment of bulimia nervosa, obsessive-compulsive disorder, panic disorder and post-traumatic stress syndrome. The latest addition is the treatment of 'social phobia' using the SSRI paroxetine (ad line: "How shy is too shy?").

Overall, this is a well-written and comprehensive account of a field that has developed rapidly in the past few decades. It was written for intelligent non-scientific readers, but experts will also find much entertaining material about the scientists who were involved in some of the major discoveries, and in the case histories. An extensive bibliography allows access to the basic scientific literature. The development of psychopharmacology has been well reviewed by David Healy in his two books published by Harvard University Press — *The Antidepressant Era* (1998) and *The Creation of Psychopharmacology* (2002) — but Barondes' new book provides a shorter and more personal account, and is an excellent read. ■

Les Iversen is in the Department of Pharmacology, University of Oxford, Mansfield Road, Oxford OX1 3QT, UK.

Science in culture

Taking it on trust

Seeing and believing the aerial photographs of Iraq.

Martin Kemp

Ever since the earliest use of illustration in a scientific text, we have been asked to take a great deal on trust. Whether recording data or placing us more literally as a surrogate witness to what something looks like, the author is proposing a contract with the reader. When an accomplished illustrator — and no one was more accomplished than Leonardo da Vinci — depicts a unicorn, we have a natural tendency to believe our eyes. Indeed, for a Renaissance spectator, a convincing unicorn was more believable than a giraffe. In our age of computer-processed images, we know that undetectable manipulation is possible, but our perceptual system is designed to collaborate with something that looks real.

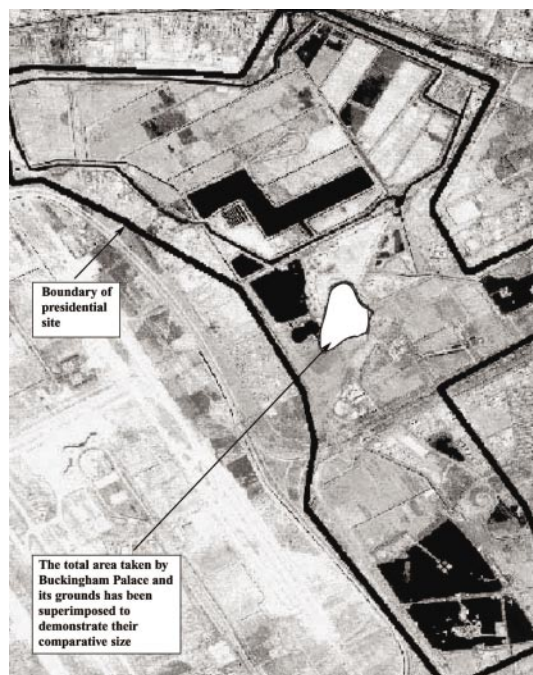
Modern science has generated a body of images of the unseeable, rather than the unseen. They demand that we trust the technological means of generation and accept the interpretative framework of expert analysers. The showing of X-rays to patients is a case in point. When a woman is shown a scan of her breast, and the surgeon points to the ominous patch, she is expected to see that "there is something there", and follow the doctor as he or she indicates that the lump is benign or malignant.

The aerial 'intelligence' photographs of Iraq used by Colin Powell at the United Nations, and subsequently published in a dossier by the British government, fall very much into this category of technological images that depend on expert intervention to transform seeing into believing.

What is captioned in the British dossier as "A photograph of a 'presidential site' or what have been called 'palaces'" provides a typical example (above). A hazy aerial view shows a grey patchwork of roads, railways, plots of land and buildings. The key points have been enhanced in black to define the boundary of the site and the complexes of buildings. To give a sense of scale, a white insert provides dramatic comparison with the "total area taken by Buckingham Palace and its grounds". Presumably a comparably scaled plan of the much larger Windsor Castle, complete with its Great Park and assorted buildings, would not have done the job as well.

The nature of the image, which we implicitly know to have been garnered through the most advanced means available, invites us to take it

as providing concrete evidence of what was happening in Iraq. 'Evidence' was an extremely heavily used word in the debates leading up to the invasion of Iraq and in the aftermath of the conquest. The models of 'information' (another widely used term) and of 'proof' used in the public debates have been derived almost exclusively from the style and terminology of modern science and technology. We are asked to rely on



the expert interpretation of such images provided by the intelligence services to the leaders who make the fateful decisions.

Those of us who use visual evidence in academic research should not automatically feel superior to politicians. When, as a sometime art historian, I draw my students into looking at something my way, I am setting up a contract of trust. I used to demonstrate to my senior students that a drawing by Michelangelo was 'obviously' a forgery — even though it was in fact an authentic drawing with an immaculate history.

Scientists who draw clarifying outlines on indistinct pictures, insert meaningful arrows on technologically generated images, or produce line transcriptions of salient features, are asking readers to believe not only that they are honest but also that their expertise really does allow them to see something on behalf of those not endowed with such insight. It is a burden of responsibility that we would do well to remember.

Martin Kemp is professor of the history of art at the University of Oxford and co-director of Wallace Kemp/Artakt.

Unnatural selection

Allison Snow

Debate about the safety of genetically engineered organisms (GEOs) has evoked intense public interest in biotechnology, and engendered varying degrees of distrust in scientists, industry and government agencies around the world. The provocative language used to describe GEOs currently in the pipeline runs the gamut from 'golden rice' to 'terminator', although the subjects of these particular monikers have not been planted commercially. In fact, much of the humanitarian promise of GEOs lies in the future, rather than in currently used products. Likewise, the technology's main hazards are probably yet to manifest themselves.

It is unfortunate that the term 'genetically engineered' has been largely overtaken by 'genetically modified' (GM) in public discourse. People have been genetically modifying crop plants, farm animals and domestic pets for centuries, most recently using 'unnatural' methods such as mutagenesis. It is the engineering aspect of transgenic technology that is both new and a cause for concern to some, as it involves the transfer of genes between sexually incompatible organisms. This approach, along with the ability to design new artificial genes and transfer them between species at will, is far more rapid and powerful than even the most sophisticated methods of conventional breeding. Thus, although the 'GM' label will continue to be used, it is important to understand the process to which this term refers.

In a stunning demonstration of engineering prowess, the plant-biotechnology firm Verdia recently designed a gene that could take the place of Monsanto's highly profitable transgene for resistance to the herbicide glyphosate. Verdia's technique involves using 'directed molecular evolution' to create new bacterial genes that can be inserted into crop plants. First, promising bacterial genes are identified by searching immense genomic databases. Mixtures of genes are then repeatedly fragmented, shuffled and reassembled under intense selection pressure in the laboratory until a useful product is obtained (such as a gene for herbicide resistance). This technology opens limitless possibilities for creating valuable, proprietary transgenes for new types of GEO. Whether the world is ready for the flood of technical achievements that biotechnology companies are eager to release is still a matter for debate.

Polarized opinions make it hard to discuss scientific questions about GEOs without delving into the reasons for the passion on both sides of this debate. To some people, the

idea of transporting genes from animals or microorganisms into food plants is repugnant and reckless, regardless of which transgenic traits are involved. It has been suggested that consumers' anxiety about transgenic organisms might be alleviated by avoiding cross-kingdom gene transfer, such as taking genes from bacteria and putting them into tomatoes, in favour of moving genes between more closely related species, such as rice and maize. But the argument that such 'within-kingdom' transgenic plants are inherently safer contradicts a basic tenet of the scientific community regarding the health and environmental effects of GEOs. For years, national science academies have argued that the products — rather than just the processes — of genetic engineering should be subject to regulation. Some would even say that the process used to acquire useful genetic variants is irrelevant; it is the phenotype that matters when evaluating possible risks.

Focusing on an organism's phenotype is desirable, given that many GEOs have truly new characteristics relative to what can be created by conventional breeding. South Korean scientists, for example, achieved an astounding 35-fold increase in the growth rates of farm-raised fish — in this case, the mud loach (*Misgurnus mizolepis*) — by fusing a native promoter to a native growth gene. In plants, the diversity of new traits that can be transferred within and between species is likely to increase dramatically as new genomics-guided strategies and high-throughput technologies are deployed. From an environmental standpoint, transgenes that confer greater crop yields might generate similar but unwanted effects in their weedy relatives through gene flow. And future types of transgenic, plant-produced pesticides may be harmful to beneficial insects and other non-target species. It is especially important to understand the environmental effects of GEOs that can proliferate freely and perhaps hybridize with other taxa.



Purple patch: transgenic fruits such as these virus-resistant plums could cut pesticide use.

Genetic engineering

Evolving techniques for redesigning organisms have enormous potential but they must be matched with equally sophisticated methods for evaluating their benefits and risks.

At the other end of the spectrum are GEOs that are phenotypically similar to conventionally bred organisms and do not establish free-living populations. Many crops have been improved using genes for disease resistance by crossing them with sexually compatible relatives — and if their relatives lack these genes, the same goal can be met by genetic engineering using genes from viruses, bacteria, plants or other organisms. As a case in point, transgenic methods are needed to develop fruit trees that resist common diseases such as apple scab and plum pox — in some cases, this could benefit consumers and the environment by reducing the need for pesticides. Genetically engineered crops that enhance yields, improve human health and make agriculture more sustainable should be encouraged, assuming that the potential health and environmental effects of these organisms are well understood.

Despite tangled regulatory systems and huge obstacles with public acceptance, harmless and beneficial uses of genetic engineering should be allowed to move forward in an ethical and scientifically informed manner. More questionable GEOs should be handled more slowly and with greater scrutiny. Genetic tinkering with self-replicating organisms requires a great deal of wisdom and caution, both from the scientists who design them and from the regulators who approve their release into the human food chain and the environment. Independent health and environmental experts can play a valuable role in guiding these decisions, but consumers will often have the final say as to which GEOs are adopted widely. ■

Allison Snow is in the Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, Ohio 43210, USA.

FURTHER READING

- Lassner, M. & Bedbrook, J. *Curr. Opin. Plant Biol.* **4**, 152–156 (2001).
 Nam, Y. K. *et al. Transgenic Res.* **10**, 353–362 (2001).
 Nielsen, K. M. *Nature Biotechnol.* **21**, 227–228 (2003).
 US National Research Council. *Environmental Effects of Transgenic Plants: The Scope and Adequacy of Regulation* (US Natl Acad. Press, Washington DC, 2002).

How to walk on water

Michael Dickinson

How the short legs of juvenile water striders propel the insects across water has perplexed researchers. It now appears that walking on water shares features with the locomotion of birds, insects and fish.

A glance at the surface of a pond reveals one of the more delightful images of summer: the shimmering ripples made by the graceful strokes of water striders. Water striders are insects that are adapted for locomotion and foraging on top of still water. Long hairy legs keep these animals afloat, but how do they glide so effortlessly across the surface? Models of water-strider locomotion have proposed that the animals move forwards by creating surface waves that carry momentum backwards. An elegant study by Hu, Chan and Bush¹, described on page 663, shows that this view is, quite literally, superficial. Like the oars of a rowing-boat, a water strider's legs create swirling vortices that carry momentum *beneath* the surface of the water. It is the rearwards motion of these vortices, and not the surface waves, that propels the animal forwards. This insight solves a paradox related to the motion of juvenile water striders, and helps to form a more cohesive picture of animal locomotion.

Much of animal locomotion distils down to a simple application of Newton's third law: to move forwards, animals must push something backwards². Just what that something is depends on the form of locomotion. Large terrestrial animals push against the solid ground, creating reaction forces in the opposite direction. The situation is a bit more complicated for swimming and flying animals, which must push against a fluid (from a physical standpoint, both air and water are fluids). As a fin or wing flaps, the fluid yields to form a pattern of swirling vortices. In some cases, the energy transmitted by an animal to the fluid in one stroke takes the form of a discrete vortex: a doughnut-shaped structure like a smoke ring. Birds³, bats⁴, insects⁵ and fish⁶ have all been shown to create a series of vortices as they move, although the precise arrangement may be complex and notoriously difficult to quantify. By carefully measuring the size, strength and velocity of the vortices generated during each stroke, it is theoretically possible to reconstruct the average force with which an animal propels itself through the air or water^{7,8}.

But what about tiny creatures such as water striders, that live between air and water? What do they push backwards to move forwards, and how do they stay afloat

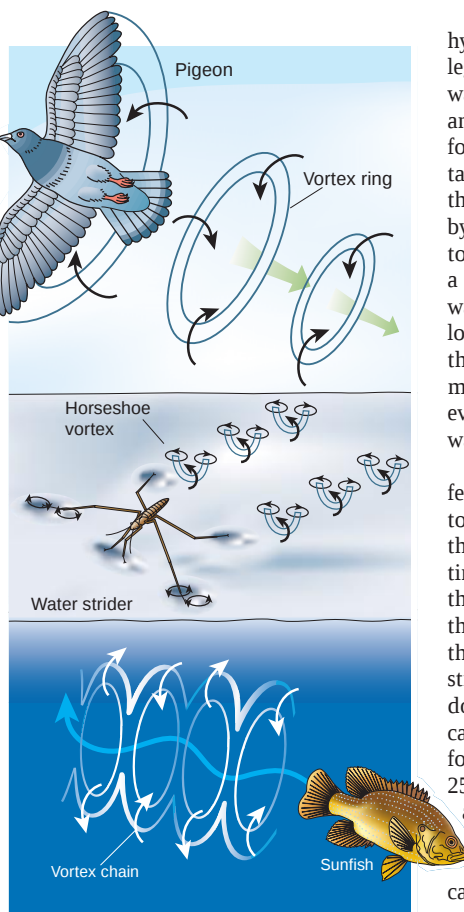


Figure 1 Locomotion in fluids. Vortices carry the momentum to move animals forwards in both air and water, but their pattern varies with the style and speed of locomotion. A slow-flying bird (pigeon) makes a single vortex ring with one downstroke. Its wake consists of a series of rings that move backwards and downwards. Each stroke of a fish's tail (sunfish) creates a vortex ring, which fuses with those of prior strokes to form a linked 'zig-zag' chain. The long legs of a water strider create tiny U-shaped vortices with their free ends attached to the water surface. In all cases, the circular flow around the vortex filament shoots fluid backwards through the centre of the ring, propelling the animal in the opposite direction.

in the first place? Resting statically on the water surface poses no great problem to a small organism⁹. Surface-tension forces at an air–water interface result from the mutual attraction of water molecules through

hydrogen bonds. As the wax-covered, hairy legs of an insect displace the water downwards, these surface-tension forces push the animal upwards. Whereas the total upward force is proportional to the perimeter of contact, which increases linearly with body size, the mass of the animal that must be supported by the tension forces increases proportionally to the cube of its body length. Thus, whereas a small water strider can easily stand atop water, larger insects need proportionally longer legs to ensure that they can rest on the surface with suitable safety. As documented by Hu *et al.*¹, this requirement for ever-longer legs limits the maximum size of a water strider to about 25 centimetres.

Understanding how water striders transfer momentum to move forwards is much tougher. One possible explanation is that they transfer momentum backwards in the tiny surface ripples created as they sweep their legs. Unlike the more familiar waves of the open ocean that are dominated by gravity, the tiny ripples generated by the legs of water striders are termed capillary waves and are dominated by surface tension. To generate a capillary wave, an insect leg (or any object, for that matter) must move faster than about 25 cm s^{-1} — the minimum speed at which any surface wave can travel. Exceeding this speed is no problem for the long legs of water-strider adults, but is beyond the capacity of the juveniles. If juvenile water striders cannot move the tips of their legs fast enough to create the surface waves, how do they propel themselves? This enigma was identified by Mark Denny⁹, and is known to water-strider enthusiasts as Denny's paradox.

The resolution of Denny's paradox, as described by Hu *et al.*, is that water striders transfer momentum not by surface ripples, but rather by vortices beneath the water surface. Unlike the circular vortices shed by fish, those created by water striders are squat affairs — a 'U-shaped' filament with two free ends attached to the water surface (Fig. 1). This peculiar shape is an elegant example of Helmholtz's laws, which govern the structure of vortex filaments. Vortex filaments cannot end abruptly within a fluid, and so must join end to end, as in a smoke ring, or attach to a wall or surface discontinuity, as in a tornado. By measuring the size and speed of the vortices formed behind the legs, the authors were able to show that the rearward

momentum created was large enough to explain the insect's forward motion, whereas the contribution of the capillary waves was much too small. So, despite their small size, the legs of water striders are analogous to the oars of a rowing-boat, which also move forwards by sending a series of vortices backwards through the fluid. This more accurate understanding of water striders fits well with the emerging picture of animal locomotion. Through their use of vortices, water striders share general features with animals flying above and swimming below them. ■

Michael Dickinson is in the Division of Engineering

and Applied Science, and the Division of Biology, California Institute of Technology, Pasadena, California 91125, USA.

e-mail: flyman@its.caltech.edu

1. Hu, D. L., Chan, B. & Bush, J. W. M. *Nature* **424**, 663–666 (2003).
2. Dickinson, M. H. *et al.* *Science* **288**, 100–106 (2000).
3. Kokshaysky, N. V. *Nature* **279**, 146–148 (1979).
4. Rayner, J. M. V. & Aldridge, H. D. J. *J. Exp. Biol.* **118**, 247–265 (1985).
5. Brodsky, A. K. *The Evolution of Insect Flight* (Oxford Univ. Press, 1994).
6. Lauder, G. V. & Drucker, E. *News Physiol. Sci.* **17**, 235–240 (2002).
7. Rayner, J. M. V. *J. Fluid Mech.* **91**, 697–730 (1979).
8. Ellington, C. P. *Phil. Trans. R. Soc. Lond. B* **305**, 115–144 (1984).
9. Denny, M. W. *Air and Water: The Biology and Physics of Life's Media* (Princeton Univ. Press, 1993).

Nanotechnology

A barrier falls

J. Tersoff

Electronic devices based on carbon nanotubes have a bright future — even more so now that a way has been found to eliminate the 'Schottky barrier' that hinders the injection of electrons into them.

Since the days of vacuum tubes, a key problem in the use of electronic devices has been getting electrons into the device from a metal connecting wire. Even the tiny transistors now being made from semiconducting carbon nanotubes share this classic problem: their performance is limited by an energy barrier that hinders electrons entering the nanotube. But now it seems that this barrier has fallen. In a paper on page 654 of this issue, Javey *et al.*¹ report the fabrication of

nanotube transistors with improved performance, achieved by reducing or eliminating the barrier that electrons must cross from the metal wire into the semiconducting nanotube.

To be sure, nanotube transistors are already performing impressively in device applications². But eliminating the barrier to electron flow is an important advance that will open the way for further improvements. The key, according to Javey *et al.*, is to use the correct combination of metal wire and

nanotube. With little or no barrier to overcome, electrons have a good chance of shooting through the nanoscale device ballistically, without being slowed by scattering. In this way, much higher currents and lower resistances are achieved than ever before — within a factor of two of the ultimate quantum limit for such a device¹.

The struggle to overcome energy barriers in electronic devices has a long history. In the venerable vacuum tube (Fig. 1), electrons face a large energy barrier to moving from the metal wire into the vacuum. The energy to overcome this vacuum barrier is provided by heating the wire until it glows brightly — an approach with obvious limitations. A similar problem plagued early devices made from familiar semiconductors such as silicon. There is an energy barrier, called the 'Schottky barrier', that electrons must overcome to get from a metal wire into the semiconductor. Modern silicon transistors circumvent the problem by replacing the metal wire with a silicon wire that is 'doped' with special impurities so that it can carry current to the device without any Schottky barrier.

Semiconducting nanotubes have the same problem: there is typically a Schottky barrier between the incoming metal wire and the nanotube. Nanotubes can be doped^{3–5}, and replacing the metal wire with a doped nanotube should also work here, just as it does for silicon⁶. But to form useful contacts in this way would require heavy doping with nanoscale spatial control — a formidable challenge. Instead, nanotube transistors have tended to rely on a geometrical advantage that they have over silicon. Because the tube is almost a one-dimensional wire, an external electric field can penetrate right to the metal–nanotube interface and reshape the barrier. If an appropriate voltage is applied, the Schottky barrier can be made so thin that electrons 'tunnel' quantum-mechanically through it. In this way, good device performance has been achieved despite any barrier², and there is room for further improvement by designing devices with this effect in mind⁵.

Nevertheless, such tunnelling always entails extra resistance: although an electron may tunnel through the thin Schottky barrier, it also has some probability of being reflected. A device's performance would be substantially improved if the Schottky barrier could be eliminated altogether, but this has not proved possible for silicon and similar semiconductors. It was anticipated that, for nanotubes, the Schottky barriers might be more easily controlled, either because the nanotube itself does not bond chemically to the metal contact, or because the nanoscale geometry makes the bonding less effective in 'pinning' the barrier height⁷. Experimental measurements support this idea⁵.

Despite these tantalizing suggestions, there had been no clear demonstration that

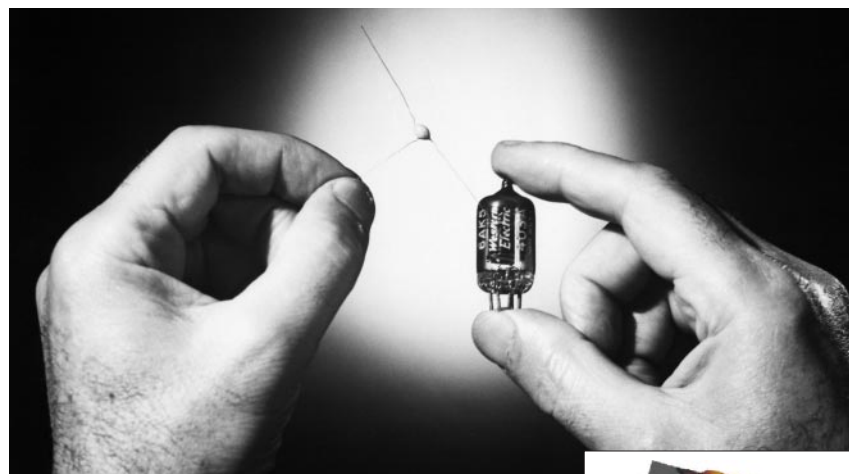
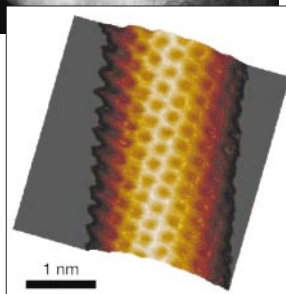


Figure 1 State of the art. Fifty years ago (when the image above was taken), the vacuum tube that had revolutionized electronics, especially in radio communications, was superseded by the transistor, a new electronic component many times smaller. Since then, transistors have shrunk to microscopic size. Now attention is turning to carbon nanotubes (an example is shown, right, in this scanning-tunnelling-microscope image; reproduced from ref. 8). Nanotubes are little more than a nanometre in diameter, but share a common problem with their electronic predecessors — how best to get electrons into the device. Javey *et al.*¹ have now found a solution.



barrier control could make a better device — until now. Using palladium for the metal connecting wire and relatively wide-diameter (about 3 nm) nanotubes, Javey *et al.*¹ have achieved this feat, eliminating the Schottky barrier and observing ballistic transmission of electrons through the device. Both choices, of palladium and of wide-diameter nanotubes, raise intriguing and as yet unanswered questions. The authors note that there is no obvious reason why palladium should give a smaller barrier than, say, platinum; palladium is distinguished primarily by its ability to stick well to carbon nanotubes. So why is its electrical behaviour different? Does palladium react chemically with the carbon tube? Does it differ from platinum or gold in barrier height, or in some other way? Can we rule out the existence of a Schottky barrier so thin as to be virtually transparent to electrons²? And can the barrier be similarly eliminated in narrower tubes, whose electrical properties are more favourable for practical devices?

In any case, Javey and colleagues' device brings us much closer to the fundamental limits of conductance; it is capable of carrying unprecedented currents at modest voltage. The evidence is compelling that this improvement comes from effectively eliminating the Schottky barrier. And although there is as yet no answer to why the barrier disappears, this work raises hope that control of the Schottky barrier will become integral to future nanotube-device technology. ■

J. Tersoff is at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA.

e-mail: tersoff@us.ibm.com

1. Javey, A., Guo, J., Wang, Q., Lundstrom, M. & Dai, H. *Nature* **424**, 654–657 (2003).
2. Wind, S. J., Appenzeller, J., Martel, R., Derycke, V. & Avouris, Ph. *Appl. Phys. Lett.* **80**, 3817–3819 (2002).
3. Zhou, C., Kong, J., Yenilmez, E. & Dai, H. *Science* **290**, 1552–1555 (2000).
4. Derycke, V., Martel, R., Appenzeller, J. & Avouris, Ph. *Nano Lett.* **1**, 453–456 (2001).
5. Heinze, S. *et al. Phys. Rev. Lett.* **89**, 106801 (2002).
6. Léonard, F. & Tersoff, J. *Phys. Rev. Lett.* **83**, 5174–5177 (1999).
7. Léonard, F. & Tersoff, J. *Phys. Rev. Lett.* **84**, 4693–4696 (2000).
8. Derycke, V., Martel, R., Radosavljevic, M., Ross, F. M. & Avouris, Ph. *Nano Lett.* **2**, 1043–1046 (2002).

Atmospheric science

African dust in Florida clouds

Owen B. Toon

Satellites and numerical models now track the intercontinental transport of airborne particles. Better knowledge of cloud physics will be necessary to gauge the effects on clouds and rainfall patterns.

If you live near a desert you expect to get dusty every now and again. Strong winds bounce sand grains across the desert surface, which in turn blast micrometre-sized particles into the air, where they blow downwind and eventually land on you. Desert dust might not seem to be a problem in a lush tropical environment such as Florida, but as our view of Earth has improved it has become clear that dust is blowing everywhere and seeping into everything. As they report in *Geophysical Research Letters*, Sassen *et al.*¹ and DeMott *et al.*² show that dust from the Sahara Desert in Africa can have an impact on clouds over Florida (Fig. 1). Sassen³ previously found that dust from Asia was affecting clouds over the western United States.

Advances in satellite remote sensing and in numerical modelling of aerosol behaviour now make possible the daily production of images and forecasts for the locations of dense dust plumes, as well as of clouds of smoke and sulphates⁴. So it was that in April 2001, awakening to visibly dirty air in Boulder, Colorado, a quick glance at the Internet⁴ showed me that local pollution was not to blame, but a dust storm in China a few days earlier. Smoke plumes from far-distant fires are also commonly observed and can now

be traced back to their sources, often on the other side of the world.

Intercontinental transport of dust has several implications, some of which have been recognized only recently. Dust scatters and absorbs sunlight, as well as infrared light radiated by the Earth, which alters the

radiation budget and is a major factor in studies of climate and climate change⁵. It is difficult to regulate air-pollution standards when they are violated by dust storms occurring halfway around the world⁶. In different contexts, wind-blown dust is a significant source of minerals for oceanic plankton, and it may be a way in which biological debris is transported over intercontinental scales. On a more local scale, a dust storm in Iran caused the collapse of a military venture to rescue the US embassy hostages in 1980, dooming the re-election prospects of President Jimmy Carter, and such storms severely affected operations in the recent conflicts in Afghanistan and Iraq.

It has long been clear that rainfall is the principal mechanism for cleansing the sky of dust, but only recently has the impact of dust on clouds been appreciated⁷. Sassen *et al.*¹ and DeMott *et al.*² now show that such effects may be widespread and not restricted to regions near deserts.

Dust may affect clouds in two ways. All water droplets start off by forming on pre-existing particles. As the number of particles increases, for instance due to a dust storm, the number of cloud droplets may increase. If there are more cloud droplets, the droplets will be smaller because the mass of condensing water is usually fixed by air motions and ambient humidity. Smaller cloud droplets make for a greater surface area, and hence brighter clouds, a consequence that was the first indirect effect of aerosols on climate to be recognized, and that has been the object of intense study over the past decade. A less well-studied phenomenon is that smaller droplets are also much less likely to collide with each other and create precipitation. Although the net flow of water through the atmosphere may be set by the rate of evaporation from the oceans, the locations of



Figure 1 Dust-up. In this image, produced with data from the SeaWiFS satellite, Saharan dust appears as a brownish haze spread across the Atlantic Ocean from Africa to the Caribbean, mingling with various cloud systems along the way. Such cross-Atlantic transport of dust is continually observed by satellites.

SEAWIFS PROJECT, NASA/GSFC, ORBIMAGE

rainfall can be shifted by such induced changes in precipitation.

By acting as nuclei for triggering ice formation, dust particles can also affect clouds by causing the water droplets to freeze at higher temperatures than expected. Pure water can become supercooled to temperatures near -40°C . But the data of Sassen *et al.*¹ and DeMott *et al.*² suggest that, in clouds over Florida, Saharan dust is causing water to freeze at temperatures between -5°C and -8°C . Ice crystals have a lower saturation vapour pressure than water droplets, so ice particles can rapidly reach large sizes in a cloud that contains both crystals and droplets. While falling, the ice particles can grow by colliding with the water droplets and induce rainfall.

Dust may thus be triggering precipitation in low-altitude clouds that otherwise would have been too warm to have produced rain, or be triggering rain at lower levels in convective clouds that otherwise would have not produced rain until reaching much higher altitudes, where it is colder. Just by acting as an abundant ice nucleus, the dust may be making high-altitude ice clouds more common. Dust may therefore inhibit precipitation by making more and smaller droplets, or enhance it by adding ice particles to warm clouds. Indeed, satellite studies over the Atlantic Ocean³ hint that both processes may occur, with dust reducing precipitation in low-level clouds and enhancing it in high-altitude clouds.

The complexity of clouds and their interactions with aerosols provide a daunting challenge in climate prediction. Global-scale climate models, upon which we rely for

climate forecasts, cannot simulate cloud physics because they lack adequate spatial resolution. Hence our climate forecasts depend upon parametrizations of clouds. Observations such as those by Sassen *et al.*¹ and DeMott *et al.*² provide the kind of insight that will spur the further development of such work. Several large field programmes are planned for the next few years, and their objectives include a better understanding of clouds and their interactions with aerosols. Ever-increasing computer resources mean that cloud-resolving models are becoming much more sophisticated and, finally, are beginning to incorporate fairly complete cloud microphysics and atmospheric dynamics on the scale of clouds. But we have a long way to go before we can provide coherent links between a dust storm in the Sahara, a cloud over Miami and the physical and chemical interactions of water on a dust grain within that cloud. ■

Owen B. Toon is in the Program in Atmospheric and Oceanic Science, and the Laboratory for Atmospheric and Space Physics, Campus Box 392, University of Colorado, Boulder, Colorado 80309-0392, USA.
e-mail: btoon@lasp.colorado.edu

1. Sassen, K., DeMott, P. J., Prospero, J. M. & Poellot, M. R. *Geophys. Res. Lett.* doi:10.1029/2003GL017371 (2003).
2. DeMott, P. J. *et al.* *Geophys. Res. Lett.* doi:10.1029/2003GL017410 (2003).
3. Sassen, K. *Geophys. Res. Lett.* doi:10.1029/2001GL014051 (2002).
4. www.nrlmry.navy.mil/aerosol
5. Sokolik, I. N. *et al.* *J. Geophys. Res.* **106**, 18015–18027 (2001).
6. Prospero, J. M. *J. Geophys. Res.* **104**, 15917–15927 (1999).
7. Rosenfeld, D., Rudich, Y. & Lahav, R. *Proc. Natl Acad. Sci. USA* **98**, 5975–5980 (2001).
8. Mahowald, N. M. & Kiehl, L. M. *Geophys. Res. Lett.* doi:10.1029/2002GL016762 (2003).

after all modifications have occurred, K-Ras, another family member, is found mainly at the plasma membrane, but Ha-Ras is present at both this and internal membranes³. Finally, active receptors in the plasma membrane can stimulate Ha-Ras on intracellular membranes⁴. These findings raised the question of how a signal can reach this Ha-Ras from surface receptors — and whether the process is biologically relevant.

Now we have answers to these questions. Bivona *et al.*² have uncovered a previously unknown signalling pathway that causes both the activation of Ha-Ras on the Golgi apparatus and its deactivation at the plasma membrane (Fig. 1). Ras proteins are 'on' when they are bound to the small molecule guanosine triphosphate (GTP), and 'off' when bound to guanosine diphosphate (GDP). The critical event in the pathway described by Bivona *et al.* is an increase in the intracellular level of calcium ions, which causes a Ras guanine-nucleotide-exchange factor (GEF), known as RasGRP1, to move to the Golgi, and a Ras GTPase-activating protein (GAP), CAPRI, to move to the plasma membrane.

GEFs activate Ras, by catalysing the replacement of bound GDP with GTP. GAPs, by contrast, turn Ras off, by enhancing the breakdown of GTP to GDP. So these findings hint at the exciting possibility that the relative levels of RasGRP1 and CAPRI determine whether Ras signals emanate primarily from the plasma membrane or from intracellular membranes (Fig. 1). The duration of the calcium increases may also affect the final outcome. Bivona *et al.* predict that, in cells displaying persistent intracellular calcium increases, Ras signalling occurs mainly from the Golgi. But when the rises are short-lived, signalling might be mostly from the plasma membrane.

The idea that the plasma membrane is not the only signalling compartment in the cell is not new^{5,6}. In particular, cellular compartments called endosomes have been proposed to have this role. The problem is that endosomes originate from the plasma membrane, making it difficult to distinguish their role in signalling from that of the plasma membrane. By contrast, Bivona and co-workers provide evidence for several biologically important roles for Golgi-activated Ras: it induces neuronal cells (PC12 cells) to differentiate and fibroblast cells to become malignant, and it makes rat intestinal epithelial cells resistant to ionizing radiation. Moreover, signalling through the T-cell antigen receptor — a key protein in the immune system — activates Ras only on the Golgi, although it is also present at the plasma membrane.

The study of signalling has traditionally involved asking 'what?' and 'when?'; it is now increasingly becoming a matter of 'where?'. But we should also ask 'why?'. Why should

Signal transduction

Life on Mars, cellularly speaking

Pier Paolo Di Fiore

A key molecular switch, known as the Ha-Ras protein, is active not only at a cell's outer membrane but also on intracellular membranes. This surprising discovery hints at unsuspected complexity in cellular signalling.

Your cells are constantly talking to one another. This dialogue relies predominantly on the emission of signals by one cell and their detection by receptors at the outer (plasma) membrane of another; the receptors then send a message inwards towards the nucleus. Inside the cell, proteins of the Ras family are key binary switches that are turned from 'off' to 'on' by these receptor-transduced incoming stimuli, and propagate signals further downstream¹. A long-held tenet is that Ras proteins work only at the plasma membrane. But we are now invited to think differently: on page 694 of this issue, Bivona and co-workers² confirm that one Ras protein, Ha-Ras, is also activated on the Golgi

apparatus inside the cell, and they show that it is not only biologically active there, but is the sole effective form of Ras in certain settings. The Golgi is involved in protein secretion, and thereby controls how signals get out of the cell. The discovery that — through Ras — it also regulates how they get in is, from a signalling viewpoint, as unexpected as life on Mars.

Previous observations had in fact already challenged the plasma-membrane-centric view of Ras. For instance, newly synthesized Ras proteins are transiently present on intracellular membranes (those of the Golgi apparatus and another cellular compartment, the endoplasmic reticulum, or ER), where the proteins are modified. Moreover,

signalling occur via intracellular membranes? Why should cells be wired in different ways to respond to the same extracellular cue? The obvious answer is to increase the complexity of cellular responses. But what does this really mean? There are several ways in which complexity could be increased by such 'compartmentalization'.

First, it might be that the 'effector' proteins downstream of Golgi Ras are different to those of plasma-membrane Ras. Current (and limited) knowledge argues against this: the known effectors of the two proteins are the same⁴. But the different locations could favour a different quantitative recruitment of effectors. Notably, plasma-membrane Ha-Ras has its full complement of possible modifications, whereas Ha-Ras on the Golgi is incompletely modified⁵. Perhaps this difference alters the protein's affinity for effectors.

Second, there is an issue of space and time. We already knew that, in PC12 cells, Ras can induce both proliferation and differentiation. Sustained activation of a major Ras effector protein, mitogen-activated protein kinase, is thought to tip the balance towards differentiation. Perhaps this sustained activation results from the stimulation of Golgi, rather than plasma-membrane, Ras. Golgi Ras is activated with delayed or sustained kinetics relative to plasma-membrane Ras^{2,4} — in other words, location (space) is translated into time — and activation of Golgi Ras triggers differentiation of PC12 cells². But how exactly is a space signal translated into a time signal? We do not know. The levels of location-specific GEFs and GAPs, and the duration of calcium ion fluxes, might be part of the answer.

Third, the presence of Ras on the Golgi hints at a variety of hitherto unsuspected roles for this protein, and for the Golgi itself. To signalling scientists, the proximity of the Golgi to the nucleus will immediately suggest that using it is a smart way of assembling signalling machineries closer to where the action really is. Researchers interested in protein secretion, meanwhile, will be intrigued by the possibility that Ras is involved in moving proteins through the Golgi. In support of this, Rho proteins — which are downstream of Ras — have already been implicated in this process⁷. In addition, Ras effectors can regulate Golgi dynamics, which in turn has been implicated in cellular processes other than protein trafficking, such as the control of cell division^{8–10}. So a role for Ras in Golgi physiology (and in the Golgi-mediated control of cell division) can be hypothesized. Finally, active Ras also occurs at the ER⁴, where several signalling pathways begin — including the unfolded-protein response, which protects the cell when the ER is under stress and unable to fold newly synthesized proteins correctly. Might this response engage Ras? And if so, does it involve the Ras on the ER?

Much remains to be done. In particular,

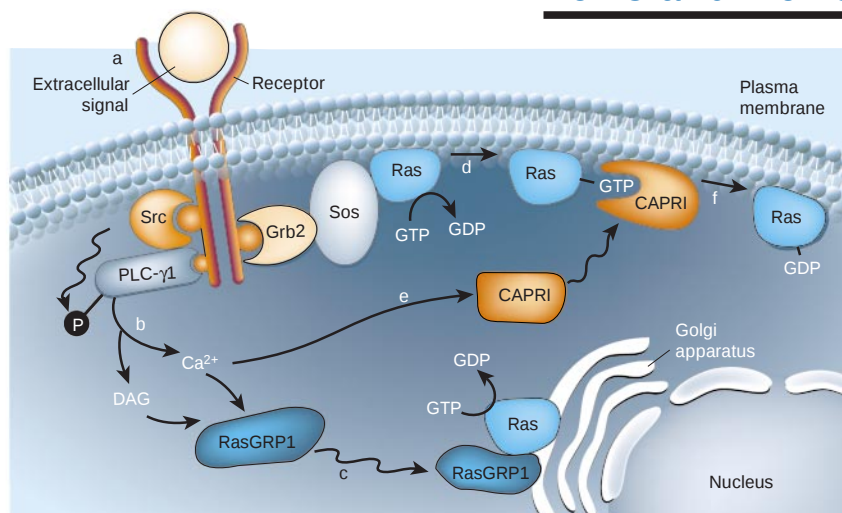


Figure 1 Switching on Ras in different cellular locations. With a few additions from the literature and reasonable assumptions, the pathway uncovered by Bivona *et al.*² can be summarized as follows. **a**, When receptors on the cell surface are activated, they recruit the enzyme Src. Simultaneously, another enzyme, phospholipase C γ 1 (PLC- γ 1), is recruited and phosphorylated by Src (represented by a circled 'P'), activating it. **b**, PLC- γ 1 leads to the generation of diacylglycerol (DAG) and an increase in the level of Ca²⁺ ions. **c**, DAG and Ca²⁺ cause a cytoplasmic protein, RasGRP1, to move to the Golgi. This protein activates Golgi-associated Ras, by catalysing the exchange of GTP for GDP. This new pathway coexists in the cell with the 'textbook' pathway, receptor $\rightarrow$ Grb2 $\rightarrow$ Sos; the latter protein activates Ras at the plasma membrane (d) in a Ca²⁺-independent way. **e**, The increase in Ca²⁺ levels also leads to the activation of CAPRI, and possibly to its movement to the plasma membrane. **f**, CAPRI inhibits Ras by stimulating its intrinsic GTPase activity, which hydrolyses GTP into GDP.

the issue of biological relevance will need further substantiation, as Bivona *et al.*² have shown that Golgi-located Ras is sufficient for certain effects on cells, but they have not proved unequivocally that it is necessary. Nevertheless, Ras has been at the centre of many exciting adventures in biology. It seems that it is set to surprise us once more.

Pier Paolo Di Fiore is at the Istituto FIRC di Oncologia Molecolare, Milan 20139, the Medical School, University of Milan, and the Istituto Europeo di Oncologia, Milan, Italy.
e-mail: difiore@ifom-firc.it

Superconductivity

Lifting the gossamer veil

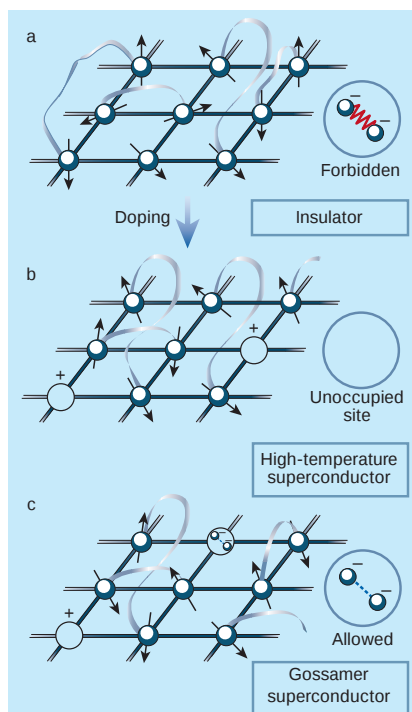
Piers Coleman

Copper oxides become superconductors at much higher temperatures than conventional metals. This transition might involve a state of 'gossamer' superconductivity, and new work shows how.

One of the outstanding mysteries in condensed-matter physics is that the most perfect conductors so far discovered — the high-temperature superconductors — are more like insulators than metals. Superconductivity, the flow without resistance of current through some materials, usually only occurs at very low temperatures. Conventional superconductivity develops in metals, but high-temperature superconductivity (at about 90 K) occurs in insulating copper oxide ceramics when

small amounts of charge are injected by chemical doping. In *Physical Review Letters*¹, Fu Chun Zhang provides a fresh perspective on this strange phenomenon by unifying alternative views of how high-temperature superconductivity is linked to an insulating state of matter. According to Zhang, an undoped copper oxide superconductor can, under certain conditions, transform from its insulating state into an extremely delicate state of 'gossamer' superconductivity².

- Schlessinger, J. *Cell* **103**, 211–225 (2000).
- Bivona, T. G. *et al.* *Nature* **424**, 694–698 (2003).
- Bivona, T. G. & Philips, M. R. *Curr. Opin. Cell Biol.* **15**, 136–142 (2003).
- Chiu, V. K. *et al.* *Nature Cell Biol.* **4**, 343–350 (2002).
- Gonzalez-Gaitan, M. *Nature Rev. Mol. Cell Biol.* **4**, 213–224 (2003).
- Sorkin, A. & Von Zastrow, M. *Nature Rev. Mol. Cell Biol.* **3**, 600–614 (2002).
- Etienne-Manneville, S. & Hall, A. *Nature* **420**, 629–635 (2002).
- Rios, R. M. & Bornens, M. *Curr. Opin. Cell Biol.* **15**, 60–66 (2003).
- Sutterlin, C., Hsu, P., Mallabiarrena, A. & Malhotra, V. *Cell* **109**, 359–369 (2002).
- Colanzi, A., Sutterlin, C. & Malhotra, V. *J. Cell Biol.* **161**, 27–32 (2003).



The quantum physicist Paul Dirac once remarked that the equations needed to understand most of physics are known, but are far too complicated to solve. Lurking in this complexity lies matter's astounding ability to develop new types of collective behaviour, such as superconductivity and magnetism. The equation that accurately describes all of this behaviour is the many-body Schrödinger equation. This involves two essential elements: the wavefunction, which is related to the probability of finding the electrons (or other particles) of a system in a given spatial configuration; and the Hamiltonian, a function that determines the energy of a system from its wavefunction. In a typical material, the number of particles tracked by the wavefunction is of the order of Avogadro's number — 6×10^{23} . This level of complexity means that understanding collective behaviour, such as that of electrons in a superconductor, depends on the creative abstraction of both the Hamiltonian and the wavefunction into a much simpler model that captures the essence of the physics.

Conventionally, superconductivity develops in a conducting metal when the electrons bind together to form 'Cooper pairs'. Each electron spins like a tiny top, and in a Cooper pair the electron spins are aligned in an antiparallel configuration. Soon after high-temperature superconductivity was discovered, it was noted^{3,4} that the insulating state from which high-temperature superconductivity arises is a special kind of insulator, called a 'Mott insulator'. The mobile electrons in a high-temperature superconductor hop from one copper atom to the next, but repel one another because of their negative

Figure 1 From insulator to superconductor.

a, Undoped copper oxides are Mott insulators. The electrons are coupled antiferromagnetically in pairs (black arrows represent spin directions) in resonating valence bonds. But the strong Coulomb repulsion between electrons prevents more than one electron occupying the site of a copper atom. Electrons cannot flow freely and hence the material is insulating. b, If a copper oxide is doped (for example by the addition of atoms with fewer electrons), the effective net removal of electrons from the insulator creates unoccupied sites into which electrons can move. The copper oxide is now a high-temperature superconductor. c, The 'gossamer' superconductor is proposed² as the link between the insulating and superconducting state. Zhang¹ suggests that the Coulomb interaction between electrons is small enough to permit double occupation of a small number of sites, freeing up some empty sites in the undoped material. The electrons now have some degree of mobility, although not as much as in the superconducting state.

charges. When these repulsive forces are large enough, the electrons are prevented from ever doubly occupying a given copper atom. In undoped copper oxide superconductors, there is precisely one mobile electron per copper atom, so the restriction on double occupancy causes the electrons to remain localized, at one electron per site (Fig. 1a). Strong magnetic forces between the localized electrons at different sites cause their spins to orient together in oppositely aligned pairs. According to the 'resonating valence bond' (RVB) model of high-temperature superconductivity^{3,4}, these pairs form a fluid that resonates between the sites in a fashion reminiscent of the Cooper pairs inside a superconductor. When charge is introduced into this insulating background of pairs (through doping with impurity atoms), the RVB state evolves directly from insulator to superconductor (Fig. 1b).

But experiment shows that the spins in the undoped insulator do not actually form the fluctuating spin fluid predicted by the RVB model. Instead, they rigidly align into antiparallel configurations to form an antiferromagnet. However, the neglect of magnetism in the theory may not be so bad. A key prediction of the RVB model — that the density of superconducting electrons, or 'superfluid density', grows linearly with chemical doping⁵ — is actually in good accord with experiments⁶. Another paradoxical property of high-temperature superconductors is that the 'gap energy' needed to remove an electron from the material actually gets larger as the superconductivity gets weaker⁷. A few years ago, Paramakanti, Randeria and Trivedi explained this feature using the RVB

wavefunction⁸, and this led to a resurgence of interest in the RVB model.

Then, about a year ago, Robert Laughlin proposed another explanation, that in fact takes the RVB approach a step further. He argued that, in their undoped state, the copper oxides should be regarded not as insulators, but as superconductors with an extremely large gap and an extremely rarefied superfluid density, which he called gossamer superconductors². In practice, the fragility of this paired state would prevent it from behaving as a bulk superconductor, but Laughlin suggested that its wavefunction might serve as a better starting point for understanding the link between Mott insulator and high-temperature superconductor.

Zhang¹ has now examined how a gossamer superconductor would develop in a model of doped copper oxide layers in which the strength of the interactions between the electrons can be tuned to become small enough to allow double occupancy of each site. The energy of repulsion between two electrons at the same site is determined by the fine details of local quantum chemistry. When this energy is large, there are no doubly occupied sites and the material is an RVB insulator. But below a certain critical value for the Coulomb interaction between the electrons, a small concentration of doubly occupied copper atoms gradually develops, leaving unfilled sites on some copper atoms. Suddenly the electrons have space in which to manoeuvre, and the RVB state transforms into a gossamer superconductor (Fig. 1c).

Zhang's new results¹ do, however, depend on approximations that require future scrutiny. One missing element is fluctuations in the phase of the superconductor⁹ — these are vital to understand how the superconducting transition temperature grows with the superfluid density. But the value of Zhang's work is that it provides a new perspective on an old problem: rather than asking how an insulator can become a high-temperature superconductor, there is now the delightful possibility that the Mott insulating ground-states of copper oxides are in fact a more delicate version of the superconductor that they become upon doping. ■

Piers Coleman is at the Center for Materials Theory, Rutgers University, 136 Frelinghuysen Road, Piscataway, New Jersey 08854-8019, USA.
e-mail: coleman@physics.rutgers.edu

1. Zhang, F. C. *Phys. Rev. Lett.* **90**, 207002 (2003).
2. Laughlin, R. B. Preprint at <http://arxiv.org/cond-mat/0209269> (2002).
3. Anderson, P. W. *Science* **235**, 1196–1198 (1987).
4. Baskaran, G., Zou, Z. & Anderson, P. W. *Solid State Commun.* **63**, 973–976 (1987).
5. Ruckenstein, A. E., Hirschfeld, P. J. & Appel, J. *Phys. Rev. B* **36**, 857–860 (1988).
6. Uemura, Y. J. *Phys. Rev. Lett.* **62**, 217–220 (1989).
7. Campuzano, J.-C. et al. *Phys. Rev. Lett.* **83**, 3709–3712 (1999).
8. Paramakanti, A., Randeria, M. & Trivedi, N. *Phys. Rev. Lett.* **87**, 217002 (2001).
9. Wang, Y. et al. *Phys. Rev. Lett.* **88**, 257003 (2002).

Neurobiology

A thorny issue

Peter H. Seeburg and Pavel Osten

A protein has been identified that makes spines grow on nerve cells. Unexpectedly, it also turns out to be part of an ion channel that is responsible for transmitting signals between neurons.

The neurons in our brain communicate with one another through specialized structures called synapses. In certain types of neuron, synapses sit on spiny protrusions. These spines are more than just structural protuberances: it seems that spine growth can be induced by stimulating neurons, indicating that changes in spine density may reflect changes in the synaptic activity — and hence the computational power — of individual neurons¹. But what are the molecular mechanisms underlying spine growth? On page 677 of this issue, Passafaro *et al.*² identify a key player in spine production — and surprisingly, it is a protein already known for its important role in synaptic function.

Synapses propagate electrical impulses from one neuron to another. In response to an electrical signal, chemical neurotransmitters are released from the axon terminal (the signal-emitting projection of neurons) of one cell and dock onto receptor proteins in the 'postsynaptic' specialization of an opposing cell. Docking leads to a brief influx of ions across the postsynaptic membrane, which takes place through a pore in the receptor itself — most neurotransmitter receptors are also ion channels. The influx of ions alters the voltage across the membrane, either depolarizing it, which creates an excitatory signal in the postsynaptic cell, or hyperpolarizing it, which leads to an inhibitory signal. The nature of the ion influx, and

hence also of the electrical signal, is determined by the type of chemical neurotransmitter: excitatory transmission is mainly produced by the chemical glutamate, whereas inhibitory transmission is mainly generated by γ -aminobutyric acid (GABA).

Spines — first described by Ramón y Cajal in 1888 as '*espinas*' (thorns) — are associated mainly with excitatory postsynaptic specializations on neurons that communicate with glutamate. Spines are connected to dendrites (the signal-receiving projections of neurons) by a neck, and they form small, semi-isolated compartments that segregate postsynaptic responses, such as increases in intracellular levels of calcium. This compartmentalization of the thousands of synaptic inputs on each excitatory neuron is widely believed to be critical for the neurons' computational power and hence for higher brain functions — a view supported by the finding that abnormalities in spine density or shape are associated with human cognitive disorders, such as Down's syndrome and fragile X syndrome³. So there is considerable impetus to identify the molecular pathways responsible for spine morphogenesis.

Many molecules have been implicated and the list comprises both signalling and structural proteins, which typically exert their functions through direct or indirect regulation of filamentous actin — the principal structural component of the spines' intracellular skeleton⁴. But how these

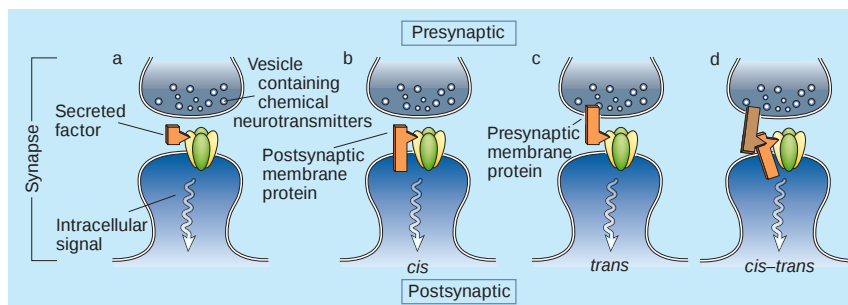


Figure 1 How to grow spines. The AMPA receptor is composed of two types of subunit: GluR2 (yellow) and either GluR1 or GluR3 (green). Interaction between the amino-terminal domain of GluR2 and another as yet unidentified protein is thought to activate a signalling cascade inside the postsynaptic neuron, triggering spine growth. Various protein–protein interactions can be envisaged between the GluR2 amino-terminal region and: a, a secreted factor released from either the presynaptic or the postsynaptic neuron; b, a postsynaptic transmembrane protein (*cis* interaction); c, a presynaptic membrane protein (*trans* interaction); d, a postsynaptic membrane protein, which in turn interacts with a presynaptic protein (*cis–trans* interaction). It seems that the molecular partner of the amino-terminal domain of GluR2 transmits the spine-inducing signal, as neither the channel activity nor the intracellular carboxy-terminal portion of GluR2 seems to be required.



100 YEARS AGO

It has occurred to me that perhaps you would care to publish the enclosed letters, and thus start some one experimenting with the radium rays in the manner suggested. Dr. Sowers is a distinguished physician of Washington, D.C., now spending a portion of his summer vacation in Baddeck, Nova Scotia.

Alexander Graham Bell

Baddeck, N.S., July 21

Dear Dr. Sowers, I understand from you that the Röntgen X-rays, and the rays emitted by radium, have been found to have a marked curative effect upon external cancers, but that the effects upon deep seated cancers have not thus far proved satisfactory. It has occurred to me that one reason for the unsatisfactory nature of these latter experiments arises from the fact that the rays have been applied externally, thus having to pass through healthy tissues of various depths in order to reach the cancerous matter. The Crookes's tube from which the Röntgen rays are emitted is, of course, too bulky to be admitted into the middle of a mass of cancer, but there is no reason why a tiny fragment of radium sealed up in a fine glass tube should not be inserted into the very heart of the cancer, thus acting on the diseased material. Would it not be worth while making experiments along this line? Yours sincerely, Alexander Graham Bell
Baddeck, N.S., July 21

Dear Dr. Bell, The suggestion which you make in regard to the application of the radium rays to the substance of deep seated cancer I regard as very valuable. If such experiments should be made, I have no doubt that they would prove successful in many cases where we now have failures. Yours sincerely, Z. T. Sowers, M. D.
Baddeck, N.S., July 21
From *Nature* 6 August 1903.

50 YEARS AGO

... the University Grants Committee does not advocate or anticipate any further expansion of the university population in the near future. It inclines to the view that there is little talent capable of benefiting by a university education which does not in fact reach the university; it is concerned rather with the redistribution inside the university which will avoid waste of public money by ensuring that... [students] are trained in ways likely to be to the public advantage. From *Nature* 1 August 1953.

molecules functionally interact to alter spine morphology remains unclear. Furthermore, the principal triggers for spine growth have, until now, been elusive.

Passafaro and colleagues² have now identified a molecule that not only increases spine density on neurons, but even makes spines grow on neurons that normally lack spines. Surprisingly, it turns out to be part of the membrane-bound ion channel that mediates the bulk of fast excitatory synaptic neurotransmission — it is the GluR2 (GluR-B) subunit of the glutamate-activated AMPA receptor.

The authors demonstrate that overproduction of GluR2 causes an increase in both the density and the size of spines in cultured hippocampal neurons (which are normally spiny), and that this effect depends on only one portion of the protein — the amino-terminal domain — which is exposed to the outside of the cell. Overproduction of either GluR2 lacking the amino-terminal portion, or a closely related AMPA-receptor subunit (GluR1), had no effect on spine growth. In contrast, a GluR1 subunit containing the amino-terminal domain of GluR2 produced spine alterations similar to those produced by the intact GluR2 subunit. Furthermore, if GluR2 production was inhibited in neurons, using a technique called RNA interference, these neurons developed fewer spines than normal. Finally, the authors showed that overproduction of GluR2 in neurons that normally have low levels of this subunit induced spines on otherwise spine-free branches of these cells. So it appears that the amino-terminal portion of GluR2 is both necessary and sufficient for the induction of spines.

Among the four AMPA-receptor subunits, GluR2 already has a special standing — it restricts calcium influx and voltage-dependent neuronal activation⁵ and it controls both the intracellular transport of newly synthesized receptors and the transport of mature receptors at the postsynaptic membrane⁶. These different functions are specified either by the particular structural configuration of GluR2's pore-forming domain within the membrane, or by determinants in the intracellular end of the protein — the carboxy-terminal domain. Passafaro *et al.* have now shown that the extracellular amino-terminal domain of GluR2 also has an important function.

How does the amino-terminal region of GluR2 work its wizardry? The authors offer three potential models of how it might activate spine growth (Fig. 1). First, it might bind a secreted factor; second, it might interact with a specific postsynaptic membrane protein (a *cis* interaction); and third, it might bind with a protein present on the presynaptic membrane (a *trans* interaction). An interaction involving both *cis* and *trans* elements could also be envisaged, as the GluR2 amino terminus could bind a postsynaptic

membrane protein that subsequently binds *in trans* to another molecule.

Such a *cis-trans* interaction has been described for the related NMDA-type receptors, which also transmit excitatory signals in response to glutamate^{7,8}. The NMDA-receptor subunit NR1 interacts, also through its amino-terminal domain, with a second membrane receptor (EphB2), which becomes clustered on the postsynaptic membrane when it binds a protein (Ephrin B) in the presynaptic membrane. During brain development, this interaction modulates the activity of the NMDA-receptor channel, leading to larger influxes of calcium ions in response to glutamate. This in turn leads to the activation of genes that affect the maturation of nerve cells, and induce the formation of spines.

So the NMDA receptor also affects spine formation, but this activity is closely coupled to its channel function. What is unique about the findings of Passafaro and colleagues² is that the effect of the GluR2 amino-terminal domain seems to be independent of ion flux through AMPA-receptor channels containing GluR2. The authors found comparable effects on spine growth with a mutant GluR2 protein that permitted high levels of ion flux (of monovalent ions as well as calcium ions) and a normal GluR2 subunit, which, when overproduced, formed channels that lacked appreciable ion flux.

Intriguingly, the effect of the GluR2 amino-terminal domain on spine morphogenesis also appears to be independent of the carboxy-terminal region. This rules out a scenario where a conformational change transduced from the amino-terminal to the carboxy-terminal domain would help to activate intracellular signalling cascades linked to spine morphogenesis. So the amino-terminal domain is likely to transmit its spine-

inducing signal to the interior of the neuron through other molecular partners (Fig. 1).

As is often the case with pioneering work, the findings of Passafaro *et al.* raise many questions. Clearly, isolation of the molecules that interact with the GluR2 amino-terminal region should be the next step in unravelling its remarkable spine-inducing ability. But even without identifying GluR2's molecular partners, it should be possible to see whether the new findings hold up in animal studies. For instance, if GluR2 really does underpin spine growth, the neurons of mice lacking the GluR2 gene⁹ should have few spines, or none at all. It would also be interesting to see whether neurons that respond to the inhibitory neurotransmitter GABA, and which normally have few spines, might become spiny in mice genetically engineered to produce GluR2 specifically in these cells¹⁰. These and other questions notwithstanding, the work of Passafaro *et al.* allows us a glimpse of how changes in synaptic activity might be connected with structural changes — an increase in the number of GluR2-containing AMPA receptors should cause parallel increases in both synaptic efficacy and spine density. ■

Peter H. Seeburg and Pavel Osten are in the Department of Molecular Neurobiology, Max-Planck Institute for Medical Research, Jahnstrasse 29, 69120 Heidelberg, Germany. e-mail: seeburg@mpimf-heidelberg.mpg.de

1. Trachtenberg, J. T. *et al.* *Nature* **420**, 788–794 (2002).
2. Passafaro, M. *et al.* *Nature* **424**, 677–681 (2003).
3. Fiala, J. C., Spacek, J. & Harris, K. M. *Brain Res. Brain Res. Rev.* **39**, 29–54 (2002).
4. Hering, H. & Sheng, M. *Nature Rev. Neurosci.* **2**, 880–888 (2001).
5. Seeburg, P. H. *Neuron* **35**, 17–20 (2002).
6. Malinow, R. & Malenka, R. C. *Annu. Rev. Neurosci.* **25**, 103–126 (2002).
7. Dalva, M. B. *et al.* *Cell* **103**, 945–956 (2000).
8. Takasu, M. A., Dalva, M. B., Zigmond, R. E. & Greenberg, M. E. *Science* **295**, 491–495 (2002).
9. Jia, Z. *et al.* *Neuron* **17**, 945–956 (1996).
10. Fuchs, E. C. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 3571–3576 (2001).

Astronomy

An elementary puzzle

Eddie Baron

A type Ia supernova has no hydrogen around it. But hydrogen gas has now been found in the vicinity of a supernova that otherwise fits the type Ia classification. Might this offer clues to the origin of these objects?

Over the past decade or so, type Ia supernovae have served as useful cosmic beacons. Using them, astronomers have measured^{1–3} the value of the Hubble constant (which determines the rate of expansion of the Universe) to an accuracy of 10%, and discovered the existence of 'dark energy'^{4,5} (which is thought to drive the acceleration of the rate of expansion). Even so, astronomers are uncertain about the origins of these objects — what progenitor system leads

to the creation of a type Ia supernova?

Supernovae are categorized according to the different chemical elements identified in the spectra of the radiation they emit. Typically, type Ia supernovae are bright, and their spectra show no evidence of hydrogen. On page 651 of this issue, Hamuy *et al.*⁶ reveal that they have detected hydrogen in the spectrum of the supernova SN2002ic, which was discovered in November last year⁷. Although in other respects SN2002ic fits the type Ia classification, the presence of hydrogen in its

vicinity sets it apart from this class, but perhaps offers a clue to its progenitor system.

Astronomers believe that type Ia supernovae are produced by the thermonuclear explosion of a white dwarf — a star that has evolved so far that its electrons are cold and densely packed, a phenomenon known as degeneracy. A white dwarf can explode when its mass grows to a limiting value, called the Chandrasekhar mass (roughly 1.4 times the mass of the Sun). To reach this limit, a white dwarf must have a companion star to feed it. In the favoured picture — the ‘single-degenerate scenario’ — this companion star is an ordinary star in a late stage of evolution, probably donating material in the form of hydrogen or helium (Fig. 1a).

An alternative theory (the ‘double-degenerate scenario’) postulates that the companion is another white dwarf, made of carbon and oxygen, which donates material as it is ripped apart by the tidal forces exerted by its more massive neighbour. In this case, no hydrogen should be seen around the eventual supernova, and until now this was true for all observed type Ia supernovae. But detailed calculations have raised doubts about whether the double-degenerate system would actually explode in a supernova, and the single-degenerate scenario is now preferred by most astronomers. Hamuy and colleagues’ detection⁶ of hydrogen around SN2002ic lends further support to the single-degenerate scenario — or perhaps rather to Hamlet’s dictum that “there are more things in heaven and earth... than are dreamt of” by conventional wisdom.

The observations by Hamuy *et al.* show that, in most respects, the spectrum of radiation from SN2002ic strongly resembles those of other well-observed type Ia supernovae. In particular, there are spectral lines that indicate the presence of silicon and sulphur, moving away from the centre of the explosion at about $10,000 \text{ km s}^{-1}$. But there is also a distinct, strong feature in the spectrum that indicates hydrogen is also present but moving much more slowly, at about $2,000 \text{ km s}^{-1}$. Such features are seen in the spectra of ‘type IIn’ supernovae, but these objects are formed when the cores of massive stars (at least eight times heavier than the Sun) collapse and explode. The supernovae of such massive stars expel dense winds, similar to the solar wind but stronger by a factor of a million or more. Their narrow spectral lines are the result of the interaction of the supernova ejecta with the circumstellar medium — the material that the massive star ejected during its lifetime (Fig. 1b). The very bright hydrogen lines seen for SN2002ic almost certainly arise in the same way. The shape of some spectral lines in SN2002ic also shows evidence of ‘toplighting’, from the combination of light from the supernova and light from the circumstellar-medium interaction⁸.

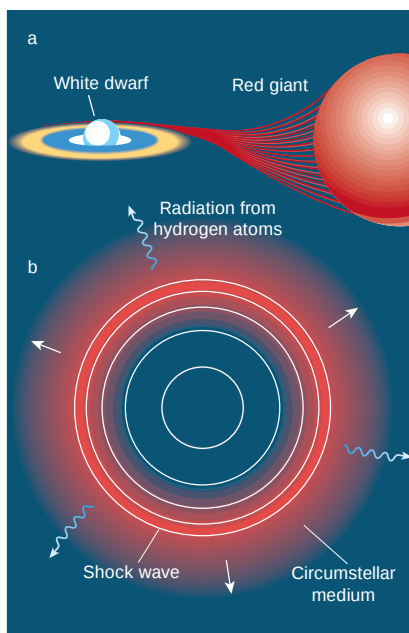


Figure 1 Type Ia, type Ix. a, In the single-degenerate scenario of supernova formation, matter from a companion star (possibly a red giant) is sucked into a white dwarf. When the mass of the white dwarf reaches the Chandrasekhar limit, it explodes as a type Ia supernova. But observations of the supernova SN2002ic by Hamuy *et al.*⁶ suggest that this particular object may be more aptly described as a ‘type Ix’ supernova, as there is evidence of hydrogen in its spectrum: b, prior to the explosion, the system may have shrugged off pulses of material, creating a hydrogen-rich envelope around it. Shock waves and light from the supernova excite atoms, including hydrogen, in the circumstellar medium, which then radiate the energy away.

The brightness of the hydrogen emission detected from SN2002ic implies that the circumstellar medium in this case is very dense, much denser than expected in the usual single-degenerate scenario⁹. If the circumstellar medium is too dense, the white dwarf won’t accrete any material; instead, it will puff up to become a red giant, and will be unlikely to increase its mass to the Chandrasekhar limit. If it does explode, the resulting supernova would appear very similar to a type II supernova. As bright objects are more easily detected than dim ones, supernovae such as SN2002ic with bright hydrogen emission lines must be rare specimens, and perhaps quite unlike a typical type Ia supernova. It might better fit a hypothetical class of supernovae¹⁰ called type Ix. Here, the white-dwarf core of a single star explodes, surrounded by an extended hydrogen envelope in a dense circumstellar medium.

To produce the narrow hydrogen line that is observed, the dense hydrogen-rich medium needs to be about 10^{10} km from the exploding white dwarf (whose radius before

explosion is only about $1,000 \text{ km}$). If that material had been ejected from the region of the white dwarf with a typical velocity of 100 km s^{-1} , it would have reached the required position in only a few years. Thus, in astronomical terms, the timing is split-second: the white dwarf must have reached the Chandrasekhar mass at practically the same time that the dense circumstellar medium was created. Such necessary coincidences always make astronomers sceptical.

Nevertheless, the data⁶ for SN2002ic seem to require a progenitor system that contains a white dwarf just ready to blow up, surrounded by a hydrogen envelope that has been thrown off just prior to explosion. It is possible that the white dwarf could have ejected a small amount of mass in a weaker, ‘nova’ explosion just before it became a supernova, but the typical masses ejected by ordinary novae are a factor of a hundred or more too small. Another possibility is that the white dwarf could have tried to explode, expelling its outer layers, and then managed to explode a year or so later, but there is no natural reason for this.

Within the single-degenerate scenario there is another possibility: the white dwarf’s companion star suddenly increased its rate of mass loss, with the result that the system puffed up into a ‘common envelope’; the white dwarf spiralled into the centre of the system, merging with the core of the companion star (M. Livio, personal communication), which increased its mass to the Chandrasekhar value. Meanwhile, frictional forces ejected the common envelope, creating the circumstellar material just in time. But, at present, SN2002ic as a modified type Ix supernova is the most appealing explanation, as it is envisaged that such stars would eject their outer layers in a series of pulses in the time before explosion.

The data from Hamuy *et al.*⁶ take us considerably closer to determining the origin of SN2002ic, but whether this will shed light on the origin of ordinary type Ia supernovae is unclear. There is no doubt that SN2002ic will keep astronomers speculating for quite a while, and they will enjoy trying to solve the puzzle.

Eddie Baron is in the Department of Physics and Astronomy, University of Oklahoma, 440 West Brooks, Norman, Oklahoma 73019-0260, USA. e-mail: baron@nhn.ou.edu

1. Parodi, B. R., Saha, A., Sandage, A. & Tammann, G. A. *Astrophys. J.* **540**, 634–651 (2000).
2. Freedman, W. L. *et al.* *Astrophys. J.* **553**, 47–72 (2001).
3. Saha, A. *et al.* *Astrophys. J.* **562**, 314–336 (2001).
4. Riess, A. *et al.* *Astron. J.* **116**, 1009–1038 (1998).
5. Perlmutter, S. *et al.* *Astrophys. J.* **517**, 565–586 (1999).
6. Hamuy, M. *et al.* *Nature* **424**, 651–654 (2003).
7. Wood-Vasey, W. M. *et al.* *IAU Circ.* 8019 (2002).
8. Branch, D., Jeffery, D., Blaylock, M. & Hatano, K. *Publ. Astron. Soc. Pacif.* **112**, 217–223 (2000).
9. Cumming, R., Lundqvist, P., Smith, L., Pettini, M. & King, D. *Mon. Not. R. Astron. Soc.* **283**, 1355–1360 (1996).
10. Iben, I. & Renzini, A. *Annu. Rev. Astron. Astrophys.* **21**, 271–342 (1983).

Physiology

Salty secrets of the Sumatran stingray

J. Exp. Biol. **206**, 2931–2940 (2003)

The stingray *Himantura signifer* spends most of its life in the fresh waters of the Batang Hari river in Sumatra, but every now and then it is thought to venture into brackish delta waters. In the lab, it can survive such salty conditions for at least two weeks. How does it cope with the shift to salt water? Wai L. Tam *et al.* provide some answers.

The authors find that, in fresh water, the stingray has higher levels of the soluble nitrogenous compound urea in its blood, compared with other freshwater species. If the external salt concentration is gradually raised, urea levels increase still further. The animals achieve this by adopting a 'waste not, want not' policy as regards ammonia: they excrete less of this waste product as salt levels increase, using it instead to synthesize more urea.

Tam *et al.* speculate that this ability may help *Himantura* to maintain its osmotic balance with the outside world and to survive in brackish waters. Its better-known cousin, *Potamotrygon motoro*, is not so versatile, and is consigned to a freshwater life in the Amazon.

Helen R. Pilcher

Physical chemistry

Stronger hydrogen bonds

J. Am. Chem. Soc. **125**, 8992–8993 (2003)

Quantum-mechanical effects strengthen hydrogen bonds, say Simone Raugei and Michael L. Klein. Their computer simulations of the simplest hydrogen-bonded liquid, hydrogen fluoride (HF), show that the intermolecular hydrogen bonds are about 4% shorter when quantum nuclear effects — the zero-point energy of the nuclei — are included, relative to 'classical' molecular-dynamics simulations.

Zero-point energy arises because quantum particles cannot sit right at the bottom of the potential well that holds them together. Intuitively, one might expect this quantum effect to make the atoms more delocalized. But the quantum simulations of Raugei and Klein show the opposite: the two fluorine atoms in a pair of hydrogen-bonded HF molecules are closer together and more localized, owing to a shortening and strengthening of the hydrogen bond.

There are as yet no experimental structural measurements on liquid HF with which to compare these simulations. They do exist for deuterated HF, but the

predicted differences between quantum and classical cases lie at the limits of experimental resolution. The researchers think that the quantum strengthening of hydrogen bonds should also apply in biological molecules such as proteins and DNA.

Philip Ball

Biomaterials

Zinc hardens worm jaws

Proc. Natl Acad. Sci. USA **100**, 9144–9149 (2003)

Hardness in nature usually comes from either calcium or silicon: think of the calcium minerals in bone and mollusc shell, or the silica exoskeletons of radiolarians. But an increasingly complex chemistry is now coming to light in nature's hard materials. Following their report of the presence of the copper mineral atacamite in the jaws of the polychaete worm *Glycera*,



Helga C. Lichtenegger and colleagues have now found zinc in the jaws of the closely related species *Nereis limbata* (pictured).

But surprisingly, whereas a definite mineral phase is seen in *Glycera* jaws, there is no sign of an inorganic zinc compound in *Nereis*. The zinc is concentrated towards the tapering tip of the jaw, and is correlated with both chlorine content and hardness. But the primary components of the jaw are glycine- and histidine-rich proteins. The researchers think that the zinc is probably bound to histidine residues in an environment like that in zinc insulin. The metal ions may serve to crosslink the protein matrix.

Philip Ball

Cell biology

Channel taps magnesium

Cell **114**, 191–200 (2003)

According to Carsten Schmitz and colleagues, a mystery mammalian ion channel called TRPM7 — which seems to be active in almost every tissue — controls the entry of magnesium ions into cells.

The channel is interesting because it is one of only two known to incorporate a

kinase domain, a structural region that modifies the activity of other proteins by adding phosphate groups to them. TRMP7's brother, TRMP6, regulates the absorption of magnesium by the kidney; people with defective TRMP6 suffer from a rare disease called primary hypomagnesaemia that is treated with mineral supplements.

Schmitz *et al.* now show that altering the structure of TRMP7, by mutating the kinase domain, increases its tendency to open in response to magnesium. This contradicts previous suggestions that the channel needs a fully functional kinase domain in order to unbolt. Moreover, TRMP7 is apparently important for letting magnesium in. Mammalian cells without a working TRMP7 stop dividing or die. But boosting the outside concentration of magnesium rescues the cells.

The authors suggest that while TRMP6 regulates the body's overall level of magnesium, TRMP7 controls the concentration in individual cells. The rush of magnesium into the cell may alter the activity of TRMP7's kinase domain, which then influences other cellular proteins.

Helen Pearson

Oceanography

Bloom in cyclone

Geophys. Res. Lett. doi: 10.1029/2003GL017141 (2003)

For three days in July 2000, cyclone Kai-Tak whipped across the South China Sea. I. Lin and colleagues have made the most of a combination of satellite views of this tropical storm and its effects, and in their latest paper have documented the resulting bloom of phytoplankton. They calculate that the bloom constituted a 10-fold increase in growth, or primary production, over normal conditions.

This is not unexpected, as storms are known to stir marine waters from depth and bring nutrients to the sunlit upper zone where phytoplankton can use them in photosynthesis. But tropical cyclones are especially unpredictable beasts, and making measurements from ships or moored arrays would be an ineffective and hazardous business. Hence the virtue of satellite data — in this case from three different sensors, which the authors have used to inform models of physical ocean mixing and primary production.

The further calculations of Lin *et al.* produce a bigger picture. Kai-Tak was of only moderate force, but they estimate that this storm alone was responsible for 2–4% of new annual production in the South China Sea. Given cyclone incidence, they believe that the overall figure could amount to 20–30%.

Tim Lincoln

Antimatter matters

John Ellis

Matter dominates antimatter, at least in our corner of the Universe. Part of the explanation could be an imbalance between the two at the level of fundamental interactions, encapsulated in the phenomenon of CP violation.

Why is the known Universe made of matter and not antimatter, if, as we believe, both were created in equal quantities in the Big Bang? An asymmetry in the behaviour of matter and antimatter at the particle level may be part of the answer. Experiments studying the decays of particles called *B* mesons are casting new light on this asymmetry. Although some results are puzzling, most agree with the predictions of a model, devised by Makoto Kobayashi and Toshihide Maskawa, of an effect known as CP violation.

The balance between matter and antimatter has been an issue since 1928, when Paul Dirac pointed out that the existence of antimatter was an unavoidable consequence of combining the theories of relativity and quantum mechanics¹. But where was the antimatter? Initially, there was speculation that the proton might be the antiparticle of the electron, despite its different mass and manifestly different properties. Once the true anti-electron — the positron — was detected in cosmic rays by Carl Anderson in 1932, serious business could begin.

A developing problem

In relativistic quantum field theory — our powerful representation of the quantum world — there are three discrete ‘symmetries’ by which particles, antiparticles and their interactions can be characterized. These symmetries correspond to transformations under which particles and their interactions may (or may not) remain unchanged (Fig. 1). The first of these is charge conjugation, commonly denoted by *C*, which does not affect the external properties of a particle such as its spin, but which reverses its internal properties such as its electric charge, changing it into its antiparticle. The second is parity or mirror symmetry, denoted by *P*, which does not affect the internal properties of a particle, but which does reverse its spin. The third transformation is time reversal, *T*, which reverses the direction of time’s arrow.

Fifty years ago, theorists proved that the combination of the three transformations, known as CPT, must be a symmetry of relativistic quantum field theory^{2–4}. In other words, under the combined CPT transformation, particle behaviour is unchanged: matter particles should behave the same as their antiparticles, if viewed in a mirror

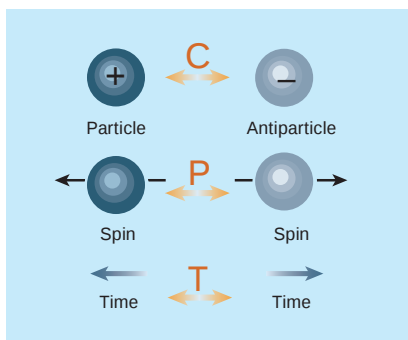


Figure 1 CPT transformations. These transformations are symmetries of quantum field theory. Applying the *C*, or charge, transformation changes a particle's internal properties, including its charge: a particle is swapped for its antiparticle and vice versa. The parity, *P*, transformation reverses the momentum of a particle and the direction in which it spins. Under the *T* transformation, the direction of time is reversed. These mathematical constructs are useful in characterizing particle behaviour — under any or all of these transformations, does the physics of particle behaviour remain unchanged?

with their internal clocks reversed. But, in 1957, some particle interactions were shown to violate each of mirror symmetry, *P*, and charge conjugation, *C*. First, it was found that radioactive nuclei tend to emit particles in directions that are related to the nuclear spins; then other interactions were found to depend on the spins of the participating particles and antiparticles; most glaringly, neutrinos and anti-neutrinos were found to spin in opposite directions.

After the initial shock of this *C* and *P* violation, physicists were relieved that at least the combination, *CP*, of these transformations seemed to be a symmetry that worked. But not for long. In 1964, *CP* symmetry was found to be broken in the decays of the subatomic particles known as *K* mesons or kaons (Fig. 2, overleaf, shows where mesons figure in the overall scheme of fundamental particles). Once in roughly every 500 times, a kaon would decay in a ‘forbidden’ way, violating *CP* symmetry (Box 1, overleaf). Fortunately, this did not conflict with any basic principles of quantum field theory⁵, but how could this *CP* violation be explained?

The shock waves were still rumbling when, in 1967, Andrei Sakharov published a remarkably perceptive paper⁶: he pointed

out that the violation of *C* and *CP* symmetries seen at this microphysical level might be responsible for the macrophysical dominance of matter over antimatter in the Universe today. In addition to breaking the *C* and *CP* symmetries, Sakharov's mechanism also required a departure from equilibrium in the early Universe and the existence of interactions that violated one of the conservation laws governing particle interactions, so-called baryon-number conservation. Unheard of at the time, violation of baryon number became popular among theorists some years later with the advent of grand unified theories, which attempt to mesh all four fundamental forces of nature in a unified framework. In 1976, Gerard 't Hooft⁷ pointed out that baryon-number violation (in combination with violation of another usually conserved quantity called the lepton number) was an inevitable part of the fledgling standard model of particle physics (which currently encompasses three of the four forces). So baryon-number violation was nothing to panic about, and Sakharov had perhaps found a *raison d'être* for *CP* violation. But still the physics of it was unexplained; and was this effect substantial enough to tip the scales in favour of matter in the early Universe?

Problem solved?

After a few years' digestion, theorists realized that *CP* violation was, after all, a natural feature of particle interactions that involved the weak nuclear force — one of the four fundamental forces. In 1973, Kobayashi and Maskawa⁸ pointed out that *CP* violation would appear in the standard model if it was extended to include six quarks (Fig. 2). At the time, only three quarks were known — called ‘up’, ‘down’ and ‘strange’ — and a fourth, the ‘charm’ quark, had been postulated, but was not seen in experiments until 1974. The suggestion of Kobayashi and Maskawa required that two more quarks should exist. Their mechanism for *CP* violation hinges on the fact that weak nuclear forces involve mixtures of different quarks with different coupling strengths. The interactions between the six types of quark are specified by nine complex numbers that can be arranged in a 3×3 matrix — the Kobayashi–Maskawa matrix (Fig. 3, page 633). If there were only four quarks, a 2×2 matrix of real numbers would be enough. But in the 3×3 version

Fundamental particles								Mesons			
Leptons/Anti-leptons				Quarks/Anti-quarks							
e^- Electron	e^+ Positron	ν_e Electron neutrino	$\bar{\nu}_e$ Electron anti-neutrino	u Up	$\bar{u}$ Anti-up	d Down	$\bar{d}$ Anti-down			$\pi^+_{u\bar{d}}$	$\pi^-_{d\bar{u}}$
μ^- Muon	μ^+ Anti-muon	ν_μ Muon neutrino	$\bar{\nu}_\mu$ Muon anti-neutrino	c Charm	$\bar{c}$ Anti-charm	s Strange	$\bar{s}$ Anti-strange			$K^-_{u\bar{s}}$	$K^0_{d\bar{s}}$
τ^- Tau	τ^+ Anti-tau	ν_τ Tau neutrino	$\bar{\nu}_\tau$ Tau anti-neutrino	t Top	$\bar{t}$ Anti-top	b Bottom	$\bar{b}$ Anti-bottom			$J/\psi_{c\bar{c}}$	$B_s_{b\bar{s}}$

Figure 2 Particles and the standard model. The fundamental particles of the standard model are leptons and quarks. From these particles, all others are made. For example, two up quarks and a down quark make a proton; adding an electron makes a hydrogen atom; anti-hydrogen is the combination of a positron and an anti-proton. Mesons are a family of two-quark particles, some examples are shown here. The ways in which these particles decay provide a useful means of studying CP violation.

for six quarks, the nine complex numbers include an imaginary 'phase', and it is this term that is the origin of CP violation in the Kobayashi–Maskawa model.

In 1977, the fifth quark — 'beauty', or more prosaically, 'bottom' — was discovered at Fermilab near Chicago. In the following years, indirect evidence, from experiments at CERN in Geneva and elsewhere, for the existence of the 'top' quark became overwhelming, although it wasn't until 1995 that it was detected directly at Fermilab. There were indeed six quarks, and Kobayashi and Maskawa were riding high. CP violation had been seen in kaon decays in 1964, but would its signature pervade all the interactions and decays mediated by weak nuclear forces, as Kobayashi and Maskawa predicted?

The first step for experimenters was to strengthen the proof of CP violation in kaon decays: the NA31 and NA48 experiments⁹ at CERN, supported by the KTeV experiment¹⁰

at Fermilab, have done so unambiguously. A new generation of experiments has now moved on to investigate in detail the decays of another family of mesons that contain bottom quarks — *B* mesons. For these particles, a high level of CP violation is expected. At the High Energy Accelerator Research Organization (KEK) in Japan and the Stanford Linear Accelerator Center (SLAC) in California, accelerators have been designed to produce a plentiful supply of *B* mesons, through specially tuned electron–positron collisions. At each facility is a detector (Belle in Japan, and BaBar in California) to pick up and study the decays of the many millions of *B* mesons created — hence these facilities are known as '*B* factories'. After a few years of data-taking, the first accurate measurements from *B* factories match, so far, the predictions of the Kobayashi–Maskawa model.

Results from these experiments are usually quoted in terms of the three angles α , β and γ

at the vertices of a triangle (Fig. 3), often called a 'unitarity triangle'. The sides of this unitarity triangle are related to elements of the Kobayashi–Maskawa matrix; the triangle would collapse to a straight line if the CP-violating phase vanished from the matrix. Different particle decays provide various handles on the different angles of the unitarity triangle, and confirm that CP is violated.

The first success of Belle and BaBar (Fig. 4, page 634), reported last year, was the measurement of the angle β (the experiments actually measure $\sin 2\beta$) with unprecedented accuracy^{11,12}, via the decays of a neutral *B* meson (B^0) to a kaon and a particle called the J/ψ ('jape-sigh') which contains a charm quark and its anti-quark. The value of the angle β agrees very well with predictions based on other weak decays, including the CP violation seen in kaon decays. The wealth of data accumulating from these and previous experiments is a resounding success for the Kobayashi–Maskawa model (Fig. 3) — many particle physicists believe that the Physics Nobel Prize Committee should be taking note.

What next?

There is much more to do. According to the Kobayashi–Maskawa model, large effects of CP violation should also be seen in many other decay modes of *B* mesons, and a number of these may now be on the threshold of observability. There are some indications of potential disagreements with the Kobayashi–Maskawa model, but the experimental statistics are still limited and dramatic conclusions are probably premature.

One interesting decay mode to watch is $B^0 \rightarrow K^0 \phi$, an analogue of the 'poster child' $B^0 \rightarrow K^0 J/\psi$, but with the charm–anti-charm meson J/ψ replaced by the strange–

Box 1 The first evidence of CP violation

The combination of a down quark and a strange anti-quark makes a K^0 meson, whereas combining a strange quark with a down anti-quark makes its antiparticle, a $\bar{K}^0$ meson. Neither of these is seen in nature: instead, they mix and appear as two other particles, one of which decays much faster than the other; hence these are called 'K-short' (K_S) and 'K-long' (K_L), respectively. If there were no CP violation, the K_S state would be the symmetric combination of K^0 and $\bar{K}^0$ (approximately $K^0 + \bar{K}^0$), and the K_L state would be the anti-symmetric combination ($K^0 - \bar{K}^0$). Furthermore, K_S would decay only into two pions (π) and K_L would decay only into three pions.

In 1964, Cronin and Fitch, with their collaborators Christenson and Turlay⁵, set up a very sensitive spectrometer (pictured) at the Brookhaven National Laboratory in New York to examine the decays of K_L particles nearly 20 m from their point of production — at that distance, short-lived K_S particles would not have survived. Cronin and Fitch were able to distinguish clearly between two-pion and other decays, and discovered that, one time in every 500, K_L particles decay into two, not three, pions. This can happen only if CP symmetry is violated.

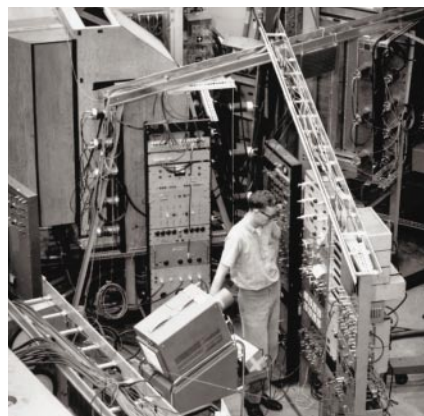
Subsequent experiments confirmed the effect and showed that CP violation arises mainly in the

mixing between the K^0 and $\bar{K}^0$: the K_L is not exactly the anti-symmetric combination of these particles, and the K_S is not exactly the symmetric combination, after all. The observed amount of mixing is consistent with the Kobayashi–Maskawa model, although it could have had other origins.

However, in 1976, it was shown³⁰ that, according to the Kobayashi–Maskawa model, effects of CP violation should also be seen

in the amplitudes for K_S and K_L decays into pairs of neutral pions and pairs of charged pions. Twenty years later, this was confirmed by experiments at CERN and Fermilab.

J.E.



BNL/SPL

anti-strange meson ϕ . According to the Kobayashi–Maskawa model, this decay should show the same CP-violating asymmetry as is seen in the J/ψ channel. Preliminary measurements^{13,14} indeed suggest a large value — but negative, rather than the expected positive value. This tentative result should be treated with caution: the statistical error in the measurement is still very large (more data are needed), and other analogous decay modes (such as $B^0 \rightarrow K^0 K^+ K^-$) do not support it. It would be prudent, therefore, to wait and see.

Caution has not prevented many theorists from speculating on what might cause a difference between the $B^0 \rightarrow K^0 J/\psi$ and $B^0 \rightarrow K^0 \phi$ asymmetries. One suggestion is that ‘supersymmetry’ is involved. This theory predicts that all the known elementary particles should be ‘shadowed’ by partners with many of the same properties, but with spins differing by half a unit. It is often thought that the interactions of these supersymmetric particles would mirror faithfully those of the standard-model particles, perhaps with the same CP-violating phase as in the Kobayashi–Maskawa model and with no other sources of CP violation. But this is probably an oversimplification; supersymmetry could accommodate dozens more CP-violating phases¹⁵. Several supersymmetrists have pointed out gleefully that decays such as $B^0 \rightarrow K^0 \phi$ might well be the most promising place to look for vestiges of supersymmetry (see, for example, ref. 16). This interpretation must contend with the absence of support from other B^0 decay modes to particles containing strange quarks, and there are anyway other interpretations. So let us indeed wait and see.

Another interesting decay mode to watch is $B^0 \rightarrow \pi^+ \pi^-$, the B -meson version of the kaon decay in which CP violation was originally discovered, and which measures a different angle of the unitarity triangle. The CP-violating asymmetry here is again expected to be large, but just how large is not known. We do not yet know enough about the underlying Kobayashi–Maskawa model parameters, and the strong nuclear interactions of the quarks involved are not easy to calculate. The BaBar experiment has not yet seen a statistically significant asymmetry¹⁷, but the Belle experiment has reported a surprisingly large and quite significant effect¹⁸. Naively combining the (rather) high Belle asymmetry with the (rather) low BaBar asymmetry would give a likely value that is comfortably consistent with the Kobayashi–Maskawa model, but it is too soon to draw such a tranquillizing conclusion.

We may not need to wait very long for answers. The B factories at KEK and SLAC are now in full operation, producing B mesons at rates that already exceed design values. The production rates are still increasing, and there are plenty of ideas for how to increase them further. The latest results will be announced on the conference circuit this summer, and we can expect that, within a few years, these experiments will have clarified the CP situation in all of the most important B^0 decay modes.

Beyond the B factories

Other experiments could also cross-check — or invalidate — the Kobayashi–Maskawa model. For example, there are decays of

K and B mesons so rare that even the B factories will have insufficient data to detect them. Some of these should be seen in other experiments looking at proton collisions: first by the general-purpose CDF and D0 experiments in operation at Fermilab; later by dedicated experiments, BTeV¹⁹ at Fermilab, and LHCb²⁰ at CERN’s Large Hadron Collider (LHC).

As well as having many more B mesons to look at, these proton-collider experiments have the advantage that they can measure large numbers of decays of mesons made out of bottom and strange quarks, the so-called B_s mesons (Fig. 2). The properties of these particles with ‘strange beauty’ can largely be predicted in the Kobayashi–Maskawa model. Checking the predictions for these particles could be the crowning glory — or the death knell — of the model.

But there are other avenues to explore. It is generally agreed that the CP violation in the Kobayashi–Maskawa model, arising from six quarks in the standard model, cannot by itself fulfil Sakharov’s mechanism. Perhaps there are other sources of CP violation, so that, taken together with the six-quark source, the effect is sufficient to generate the matter–antimatter asymmetry of the Universe. Theorists confidently expect that, to complete our understanding of the particle world, there are many extensions of the standard model to be made, and each of these introduces possible new mechanisms for CP violation.

One possible extension already mentioned is supersymmetry. If the theory is right, a new family of particles should be observable at the LHC. There could be dozens

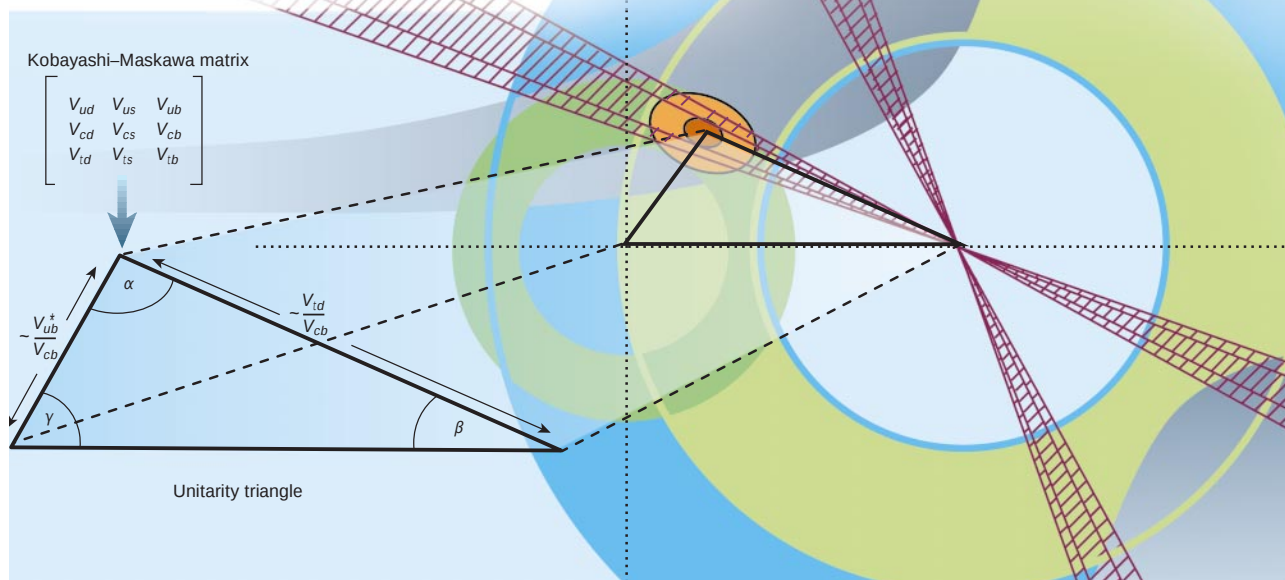


Figure 3 The Kobayashi–Maskawa matrix and the unitarity triangle. The terms of the Kobayashi–Maskawa matrix are complex numbers representing the coupling strengths between the six different quarks of the standard model. Their relation to each other is often represented in the form of a triangle — the unitarity triangle — whose dimensions are proportional to combinations of the matrix terms (V_{ub} , shown here, is the complex conjugate of V_{ub}). Drawing the triangle defines three angles, and these are the quantities being measured in the current round of experiments at the B factories. Data from many sources, fed into a single plot³¹ (right), already constrain the values of the three angles quite tightly. The B factories will produce more accurate results, to be followed up in even finer detail in the next generation of experiments.

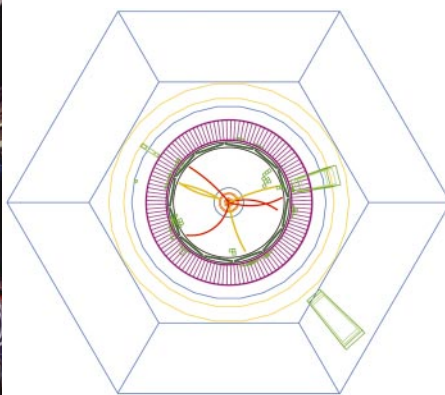
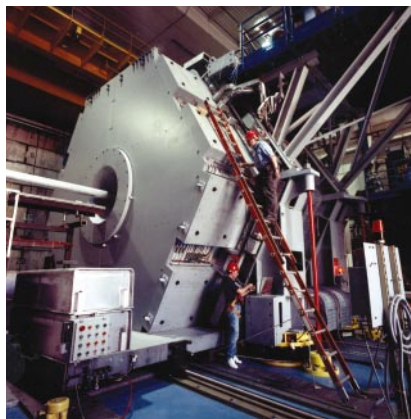


Figure 4 The BaBar detector. In operation since 1999, BaBar (pictured left, under construction) has analysed the debris of more than 90 million electron–positron collisions. From the energy released by the annihilation of electron and positron in the centre of the detector, *B* mesons are created, which typically decay into the spray of particles seen here (right), including muons, kaons and pions.

of new CP-violating phases in their interactions, some of which might have been able to generate the Universe's matter–antimatter asymmetry around the epoch in the early Universe when the masses of the elementary particles were generated, through the Higgs mechanism. There is no direct evidence yet for the related Higgs boson, but this is the principal prey of present Fermilab and future LHC²¹ experiments. Is this what lies behind the Sakharov mechanism? A key test could be the observation of CP-violating effects in the production and decays of Higgs bosons, first at the LHC, then perhaps to be followed up with more detailed measurements at a future, very-high-energy, electron–positron collider — if the Higgs boson indeed exists.

Matter from neutrinos?

Another exciting arena for CP violation is neutrino physics, which is the one place where physics beyond the standard model has already appeared, in the form of oscillations between the three species of neutrino (Fig. 2). These oscillations reflect mixings of the neutrinos that involve a CP-violating phase (first discussed by Maki, Nakagawa and Sakata²²), in close analogy with the Kobayashi–Maskawa phase in the mixing of quarks. Neutrino oscillations have been uncovered in observations of neutrinos reaching Earth from the Sun or created in the planet's atmosphere. More will be learned in a new round of experiments currently being prepared in Japan using a man-made beam of neutrinos and the SuperKamiokande detector (reincarnated, after most of its sensitive components imploded in 2001).

The long-term goal of these experiments and their successors will be to observe either a CP-violating asymmetry between the oscillations of neutrinos and of anti-neutrinos, or a T-violating asymmetry (for example, if the oscillation probability of an electron neutrino to a muon neutrino does not match that of the reverse process). Ideal for the

task²³ would be a 'neutrino factory', using a pure neutrino beam produced by the decay of other particles, either muons²⁴ or possibly radioactive nuclei²⁵.

CP violation in neutrinos could then also be part of the Sakharov mechanism²⁶, albeit in a roundabout way. In most models for neutrino masses, the three light neutrinos we know are accompanied by as yet unseen, heavy partners (weighing at least 10^{10} giga-electronvolts; for comparison, a proton weighs 1 giga-electronvolt). In a scheme known as 'leptogenesis', the decays of these heavy beasts could generate an asymmetry in the early Universe between the particle family called leptons (which includes electrons) and the anti-leptons (including positrons); 't Hooft effects⁷ would subsequently convert part of this into an asymmetry between quarks and anti-quarks, and hence between matter and antimatter. In principle, the ingredients of leptogenesis can be tested in laboratory experiments (for example, by looking for neutrinoless double beta decay, or an electric dipole moment of the electron or the muon), but it is an ambitious and long-term programme²⁷.

Little ado about CPT?

Throughout this discussion of the matter–antimatter asymmetry, CP violation has been emphasized and not much has been said about CPT violation. In fact, a number of experiments have already constrained this possibility severely. For example, we know that the masses of kaons and anti-kaons are the same to better than one part in 10^{18} — leaving only a small window for the violation of CPT.

But CERN has recently embarked on an experimental programme that will approach CPT violation from another direction: to look for any differences between the structure (as seen in the spectra of emitted radiation) of hydrogen and anti-hydrogen, down to a level of one part in 10^{12} or 10^{15} . Admittedly,

we theorists do not really expect that CPT violation will show up in these experiments, as that would mean abandoning the sacred principles of quantum field theory that have worked so well for more than 50 years — but we have been wrong before. Two experiments, ATHENA²⁸ and ATRAP²⁹, have already reported the creation of 'industrial quantities' of anti-hydrogen atoms, around 100,000 each, and the race is now on to measure their spectra.

So what?

The matter–antimatter asymmetry is one of the subtlest properties of the microphysical world. Although its discovery in experiment was unexpected, theorists have learned to live with it, and even to love it. The Kobayashi–Maskawa model of CP violation reigns supreme in the quark sector, despite the efforts of theorists and experimenters alike to knock it off its pedestal. But neutrino physics and supersymmetry (if it exists) bid fair to provide additional sources of CP violation, and may complete the link to one of the most profound questions in macro-physics: the origin of the matter dominating our Universe. Studies of CP violation may not only reveal to us voyeuristic secrets of particles' intimate behaviour, but also cast light on the origins of our existence. ■

John Ellis is in the Theory Division, CERN, CH-1211 Geneva 23, Switzerland.

e-mail: john.ellis@cern.ch

1. Dirac, P. A. M. *Proc. R. Soc. Lond. A* **117**, 610–624 (1928).
2. Pauli, W. in *Niels Bohr and the Development of Physics* (eds Pauli, W., Rosenfeld, L. & Weisskopf, V.) 30–51 (McGraw-Hill, New York, 1955).
3. Lüders, G. *Ann. Phys. (NY)* **2**, 1–15 (1957).
4. Jost, R. *Helv. Phys. Acta* **31**, 263–272 (1958).
5. Christenson, J. H., Cronin, J. W., Fitch, V. L. & Turlay, R. *Phys. Rev. Lett.* **13**, 138–140 (1964).
6. Sakharov, A. *Pisma Zh. Eksp. Teor. Fiz.* **5**, 32–35 (1967).
7. 't Hooft, G. *Phys. Rev. Lett.* **37**, 8–11 (1976).
8. Kobayashi, M. & Maskawa, T. *Prog. Theor. Phys.* **49**, 652–657 (1973).
9. Batley, J. R. *et al. Phys. Lett. B* **544**, 97–112 (2002).
10. Alavi-Harati, A. *et al. Phys. Rev. D* **67**, 012005 (2003).
11. Aubert, B. *et al. Phys. Rev. Lett.* **89**, 201802 (2002).
12. Abe, K. *et al. Phys. Rev. D* **66**, 071102 (2002).
13. Aubert, B. *et al.* Preprint at <http://arXiv.org/hep-ex/0207070> (2002).
14. Abe, K. *et al. Phys. Rev. D* **67**, 031102 (2003).
15. Masiero, A. & Vives, O. *New J. Phys.* **4**, 1–19 (2002).
16. Ciuchini, M., Franco, E., Masiero, A. & Silvestrini, L. *Phys. Rev. D* **67**, 075016 (2003).
17. Aubert, B. *et al. Phys. Rev. Lett.* **89**, 281802 (2002).
18. Abe, K. *et al. Phys. Rev. D* **68**, 012001 (2003).
19. www-btev.fnal.gov/btev.html
20. http://lhcb.web.cern.ch/lhcb
21. Carena, M., Ellis, J. R., Mrenna, S., Pilaftsis, A. & Wagner, C. E. Preprint at <http://arXiv.org/hep-ph/0211467> (2003).
22. Maki, Z., Nakagawa, M. & Sakata, S. *Prog. Theor. Phys.* **28**, 870–880 (1962).
23. Cervera, A. *et al. Nucl. Phys. B* **579**, 17–55 (2000); Erratum **593**, 731 (2001).
24. Apollonio, M. *et al.* Preprint at <http://arXiv.org/hep-ph/0210192> (2002).
25. Zucchelli, P. *Phys. Lett. B* **532**, 166–172 (2002).
26. Fukugita, M. & Yanagida, T. *Phys. Lett. B* **174**, 45–47 (1986).
27. Ellis, J. R. & Raidal, M. *Nucl. Phys. B* **643**, 229–246 (2002).
28. Amoretti, M. *et al. Nature* **419**, 456–459 (2002).
29. Gabrielse, G. *et al. Phys. Rev. Lett.* **89**, 213401 (2002).
30. Ellis, J. R., Gaillard, M. K. & Nanopoulos, D. V. *Nucl. Phys. B* **109**, 213–243 (1976).
31. http://ckmfitter.in2p3.fr

A cloned horse born to its dam twin

A birth announcement calls for a rethink on the immunological demands of pregnancy.

Several animal species, including sheep, mice, cattle, goats, rabbits, cats, pigs and, more recently, mules¹ have been reproduced by somatic cell cloning, with the offspring being a genetic copy of the animal donor of the nuclear material used for transfer into an enucleated oocyte. Here we use this technology to clone an adult horse and show that it is possible to establish a viable, full-term pregnancy in which the surrogate mother is also the nuclear donor. The cloned offspring is therefore genetically identical to the mare who carried it, challenging the idea that maternal immunological recognition of fetal antigens influences the well-being of the fetus and the outcome of the pregnancy.

We derived equine fibroblast cell lines from skin biopsies taken from one Arabian thoroughbred male and one Haflinger female. Oocytes were obtained by *in vitro* maturation of follicular oocytes recovered from equine ovaries retrieved from abattoirs. After removal of the nucleus and zona material, embryos were reconstructed by cell fusion with the fibroblasts. Only fused embryos (97%) were activated and cultured *in vitro* to the blastocyst stage; these were then transferred into recipient mares (see supplementary information).

We cultured 513 and 328 successfully fused reconstructed embryos from the male and the female cell lines, respectively. In total, 22 blastocysts developed, 8 from the male and 14 from the female cell line. Of these, eight blastocysts from the male cell line and nine from the female line were transferred non-surgically into nine recipients. Two of the cloned embryos that were reconstructed with female cells were transferred into the Haflinger mare from which the cell line was derived.

Overall, four single pregnancies were diagnosed by ultrasonography after 21 days of gestation. Two were lost soon afterwards and one aborted at 187 days of gestation; the fourth, however, was carried to term (336 days) and a healthy female foal was born naturally on 28 May 2003 (birth weight, 36 kg). Remarkably, this cloned foal was born from the same Haflinger mare who was the original cell-line donor (Fig. 1).

To confirm the genetic identity of the foal and her mother, we analysed DNA from blood leukocytes from both animals and from the placenta. Comparison of 12 equine-specific microsatellite loci (Stock-Marks kit, Perkin-Elmer, Applied Biosystems) confirmed that DNA from the foal and from the placenta was identical to that of the recipient mare.

Our cloning procedure was relatively efficient, as one live foal was produced from four pregnancies, although there was high developmental failure from the cleavage stage to blastocyst (8 of 467 and 14 of 286 developed in the male and female cell lines, respectively) and early implantation. This success was aided by advances in assisted reproduction in the horse², particularly at the oocyte-activation stage, when protein synthesis and phosphorylation must both be inhibited³, and in the refinement of the zona-free manipulation technique^{4,5}.

The remarkable birth of a live foal from its genetically identical recipient is at odds with the idea that the maintenance of gestation depends on immunological recognition of the pregnancy by the mother, based on evidence that abortion can result from inadequate recognition of fetal antigens⁶.

Furthermore, our result adds the horse to the list of mammals that have so far been successfully cloned from an adult somatic cell. In principle, cloning could enable gelding champions to contribute their genotype to future generations, as well as opening up an opportunity to verify the reproducibility of traits such as character and sporting performance.

Cesare Galli, Irina Lagutina, Gabriella Crotti, Silvia Colleoni, Paola Turini, Nunzia Ponderato, Roberto Duchi, Giovanna Lazzari

Laboratorio di Tecnologie della Riproduzione, CIZ srl, Istituto Sperimentale Italiano Lazzaro Spallanzani, 26100 Cremona, Italy



Figure 1 The cloned foal, named Prometea, with its genetically identical mother seven days after birth. The cells used for cloning were derived from a skin biopsy of the mare. The enucleated oocytes were obtained by *in vitro* maturation and removal of the nucleus from oocytes recovered from horse ovaries from an abattoir. After embryo reconstruction by micromanipulation and cell fusion, the cleaved embryos were cultured *in vitro* to the blastocyst stage. Two of the derived embryos were transferred into the mare donor of the cell line. A single pregnancy was established and was carried to term with the birth of a female foal who is genetically identical to her surrogate mother.

e-mail: cesare@galli2.191.it

1. Woods, G. L. *et al.* *Science* doi:10.1126/science.1086743 (2003).
2. Galli, C. *et al.* *Theriogenology* **58**, 705–708 (2002).
3. Lazzari, G., Mari, G. & Galli, C. *J. Reprod. Fert. Abstr. Ser.* **28**, 73 (2002).
4. Lagutina, I. *et al.* *Theriogenology* **59**, 269 (2003).
5. Wilmut, I. *Cloning and Stem Cells* **5**, 1–2 (2003).
6. Szekeres-Bartho, J. *Int. Rev. Immunol.* **21**, 471–495 (2002).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Lithium-ion batteries

Runaway risk of forming toxic compounds

Lithium-ion batteries are stabilized by an ultrathin protective film that is 10–50 nanometres thick and coats both electrodes. Here we artificially simulate the 'thermal-runaway' conditions that would arise should this coating be destroyed, which could happen in a battery large enough to overheat beyond 80 °C. We find that under these conditions the reaction of the battery electrolyte with the material of the unprotected positive electrode results in the formation of toxic fluoro-organic compounds. Although not a concern for the small units used in today's portable devices, this unexpected chemical hazard should be taken into account as larger and larger lithium-ion batteries are developed,

for example for incorporation into electric-powered vehicles.

Lithium-ion cells are now ubiquitous in consumer electronics. The negative electrode consists of lithium in graphite and the positive electrode contains lithium and a transition metal such as cobalt, nickel or manganese. However, with a lead voltage of about 4 volts, no electrolyte is thermodynamically stable towards either electrode, so the operation of these batteries relies on a combination of the ethylene carbonate and LiPF₆ (LiBF₄)-based electrolyte producing a continuous film that ensures sufficient ionic conductivity and provides electronic insulation to protect the electrodes.

Above 80 °C, thermal runaway can occur spontaneously as a result of the break-up of this protective film. Today's commercial batteries are checked by electronic monitoring of individual cells and by a self-clogging polyolefin separator to prevent overheating,

Table 1 Product yields from different lithium compounds

Run	A	Solvent	B	Organo-F yield (%)
1	LiPF ₆	EC	—	0
2	LiPF ₆	EC	LiCoO ₂	80
3	Li[(CF ₃ SO ₂) ₂ N]	EC	LiCoO ₂	0
4	LiBF ₄	EC	LiCoO ₂	80
5	LiBF ₄	EC	LiMn ₂ O ₄	80
6	LiBF ₄	EC	LiNiO ₂	80
7	LiPF ₆	EC + DMC	LiCoO ₂	35

Yields of fluoro-organics are shown for ethylene carbonate reacting with lithium compounds A and B at 240 °C for 30–60 min in different solvents. EC, ethylene carbonate; DMC, dimethyl carbonate.

and their failure rate has dropped to about 0.3 per million. The consequences of runaway reactions have so far focused mainly on the risks of fire, burns and the release of hydrogen fluoride.

The compound 2-fluoroethanol is highly toxic (the 50%-lethal dose, LD₅₀, in mice is 0.1 mg kg⁻¹) and is readily formed by nucleophilic attack of fluoride ion on ethylene carbonate¹. Although no example of electrophilic fluorination using coordination anions such as LiPF₆ is known, we investigated whether a similar reaction could occur by this route.

Ethylene carbonate and LiPF₆ alone do not form fluoro-organics, but these are generated in abundance when any of the positive-electrode materials (Li_xMO₂, where M is Co, Ni or Mn) is also present. Gas chromatography with mass spectrometry (110 mass units) indicates that the fluoroethanol ether is formed under these conditions: two moles of ethylene carbonate react with one mole of LiPF₆ to yield one mole each of (FCH₂CH₂)₂O, LiF and volatile POF₃ (boiling point, -39.7 °C), providing the driving force with the loss of two moles of CO₂.

Ethylene oxide (formed by the loss of CO₂ from ethylene carbonate and catalysed by the transition metal) is a likely transient intermediate. The compound bis-(2-fluoroethyl)-ether, which has an LD₁₀₀ in rats of 10 mg kg⁻¹ (estimated LD₅₀ of 4 mg kg⁻¹; ref. 2), represents 80% of the total fluoro-chemicals in runs 2 and 4–6, which are summarized in Table 1. The carbon anion with covalently bonded fluorine (CF₃⁻) is inert.

Our experiments are *ex situ* and relate only to the reaction of electrolyte with the material of the positive electrode. However, the conditions triggering the ring-opening fluorination of ethylene carbonate could easily occur in large batteries (weighing more than 500 g), such as those proposed for use in electric vehicles, where a short-circuit might cause the temperature in the electrolyte to rise instantly above 250 °C. Although these batteries are in principle highly desirable for saving energy and preventing pollution and for CO₂ abatement, their safety should be rigorously investigated under protected conditions. Further exploration of the chemicals evolved in runaway reactions should be undertaken with

extreme caution and only where *ad hoc* facilities are available.

Amer Hammami, Nathalie Raymond, Michel Armand

Laboratoire International sur Les Matériaux Electro-actifs CNRS-UdM, UMR 2289, Université de Montréal, PO Box 6122, Montréal, Québec H3C3J7, Canada
e-mail: michel.armand@umontreal.ca

1. Bergmann, E. D. & Shahak, I. J. *Chem. Soc. C*, 899–900 (1966).
2. Gitter, S., Blank, I. & Bergmann, E. D. K. *Ned. Akad. Wetenschap. Proc. C* 56, 427–430 (1953).

Competing financial interests: declared none.

Palaeontology

Spider-web silk from the Early Cretaceous

The use of viscid silk in aerial webs as a means to capture prey was a key innovation of araneoid spiders and has contributed largely to their ecological success¹. Here I describe a single silk thread from a spider's web that bears glue droplets and has been preserved in Lebanese amber from the Early Cretaceous period for about 130 million years. This specimen not only demonstrates the antiquity of viscid silk and of the spider superfamily Araneioidea, but is also some 90 million years older than the oldest viscid spider thread previously reported in Baltic amber from the Eocene epoch².

The silk glands that produce the gluey coating on viscid silk may be the best single character to define the Araneioidea, the superfamily that includes all ecribellate orb-weavers and the comb-footed spiders

(Theridiidae)³. A fossil spider from the Early–Middle Jurassic (which is about 190 Myr old) has been described — on the basis of characters other than its silk-producing glands — as belonging to the Araneioidea⁴, and araneoid spiders have been recorded from the Early Cretaceous^{5,6}.

Spiders are known to have been able to produce silk since the Mid-Devonian period, about 410 Myr ago⁷. It is unclear, however, how today's intricate araneoid webs with viscid silk evolved from the probably primitive silk that was used by Devonian spiders. Current ideas about the evolution of aerial webs and of viscid silk are based on the phylogeny of the spiders, which in turn is based on morphological characters — mostly genitalic ones — whose evolution is generally not directly linked to the spider's ability to build webs or to produce viscid silk.

The present specimen (C19/2, held at the Staatliches Museum für Naturkunde in Stuttgart, Germany) was collected by Dieter Schlee in 1969 from amber beds located near Jezzine in Lebanon. These deposits are dated to the Hauterivian (about 127–132 Myr) and represent the oldest known amber with insect inclusions⁸.

The specimen contains a single thread of about 4 mm in length, which shows a striking resemblance to recent araneoid spider threads, with 38 distinct fossilized glue droplets arranged at regular intervals along two sections (Fig. 1a). The thread diameter (3 µm) and the size (7–29 µm), density (22 droplets per mm), arrangement (small droplets alternating with larger ones) and shape (semi-transparent globules, slightly elongated along the thread) of most glue droplets (Fig. 1b) closely match those in recent araneoid webs⁹. Some droplets have a larger diameter of up to 85 µm (Fig. 1c), probably as a result of water uptake from the resin (araneoid glue droplets are highly hygroscopic¹⁰) and subsequent merging with neighbouring droplets.

As the specimen contains only a single thread, we cannot identify with certainty the type of web of which this thread was a part. It may have been part of an orb-web or of a gum-footed web, as built by some

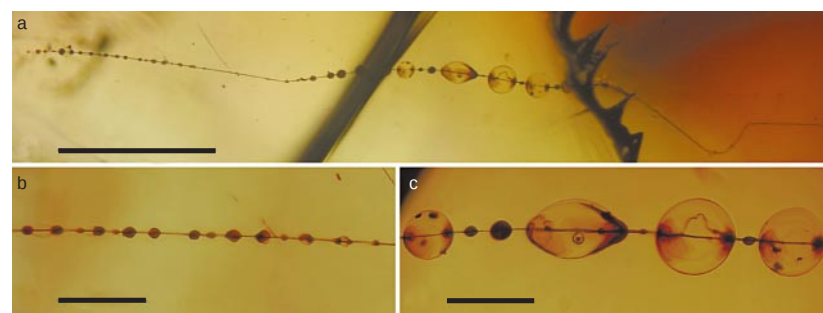


Figure 1 Fossil araneoid spider thread with glue droplets in Lebanese amber. **a**, Overall view: the thread runs at an angle of 50° to the upper surface of the amber; the left-hand end of the thread reaches the lower surface of the amber. **b**, A region of the thread showing regularly arranged glue droplets. **c**, Distended glue droplets with clearly visible core threads. Scale bars: **a**, 500 µm; **b**, **c**, 100 µm.

recent theridiid spiders. In these webs, only the gumfeet, the peripheral parts of the threads extending from a central structure to the substrate, are equipped with glue droplets¹¹.

It is likely that the present specimen was such a gumfoot, because viscid threads in other araneoid webs are always surrounded by non-sticky threads, and because gumfeet are the only type of viscid thread that is built near the bark of a tree, where the probability of becoming engulfed in resin is largest. In addition, the thread is only partly covered with glue droplets, another feature that it shares only with gumfeet.

However, the fossil thread could have been part of a web type that no longer exists today. In any case, the present specimen is direct evidence for the antiquity of viscid spider silk and provides independent evidence for the antiquity of the Araneioidea.

Samuel Zschokke

Department of Integrative Biology, Section of Conservation Biology, University of Basel, 4056 Basel, Switzerland
e-mail: samuel.zschokke@alumni.ethz.ch

1. Opell, B. D. *Funct. Ecol.* **12**, 613–624 (1998).
2. Bachofen-Echt, A. *Biologia Generalis* **10**, 179–184 (1934).
3. Coddington, J. in *Spiders: Webs, Behavior, and Evolution* (ed. Shear, W. A.) 319–363 (Stanford Univ. Press, California, 1986).
4. Eskov, K. *Neues Jb. Geol. Paläontol. Mh.* **1984**, 645–653 (1984).
5. Selden, P. A. *Nature* **340**, 711–713 (1989).
6. Penney, D. & Selden, P. A. *J. Arachnol.* **30**, 487–493 (2002).
7. Shear, W. A., Palmer, J. M., Coddington, J. A. & Bonamo, P. M. *Science* **246**, 479–481 (1989).
8. Schlee, D. & Dietrich, H.-G. *Neues Jb. Geol. Paläontol. Mh.* **1970**, 40–50 (1970).
9. Opell, B. D. *J. Arachnol.* **30**, 494–502 (2002).
10. Vollrath, F. et al. *Nature* **345**, 526–528 (1990).
11. Benjamin, S. P. & Zschokke, S. *Biol. J. Linn. Soc.* **78**, 293–305 (2003).

Competing financial interests: declared none.

Insect signalling

Components of giant hornet alarm pheromone

Up to 74 people die each year in Japan after being stung by Hymenopteran insects, with hornets (*Vespa* spp.) being among the worst offenders¹. Here we identify a volatile, multi-component alarm pheromone in the venom of the world's largest hornet, *V. mandarinia*, and use field bioassays to show that 2-pentanol is its principal active component, and that 3-methyl-1-butanol and 1-methylbutyl 3-methylbutanoate act synergistically with it. The compound 1-methylbutyl 3-methylbutanoate, which may also be a foraging-site-marking pheromone, elicits a strong defensive reaction in the sympatric prey hornet *V. simillima xanthoptera*. As these chemicals are sometimes used in food flavourings and as fragrances in cosmetics^{2,3}, it is possible that they might provoke a seemingly unwarranted hornet attack on humans.

We investigated whether the strongly defensive behaviour of *V. mandarinia* could be due to the presence of a powerful alarm pheromone in the venom by capturing the volatile components from the venom sacs of three frozen worker hornets by solid-phase micro-extraction (SPME) and analysing them by gas chromatograph mass spectrometry. These volatile components were absorbed onto SPME fibre (with a 100- μ m coating of polydimethylsiloxane), which was then inserted into the injection port of the spectrometer at 250 °C.

Three biologically active chemicals — 2-pentanol, 3-methyl-1-butanol and 1-methylbutyl 3-methylbutanoate — were identified from the venom. Analysis against synthetic enantiomers and using a chiral column revealed that the (*S*) enantiomer was predominant, with eight times less (*R*) enantiomer for 2-pentanol and 1-methylbutyl 3-methylbutanoate.

In field bioassays, the crude venom extract caused intense alarm and defensive behaviour, with worker hornets flying excitedly around the nest and rushing towards the target. Authentic (*R*), (*S*) and racemic 2-pentanol elicited the same response. A synergistic effect was observed with the addition of 1-methylbutyl 3-methylbutanoate and 3-methyl-1-butanol (Figs 1, 2). Several commercial products containing these chemicals also elicited this defensive reaction. The use of these extremely volatile chemicals as alarm signals may allow a hornet colony to resume normal behaviour rapidly — that is, very soon after an intruder has been driven away and pheromone release ceases.

We also identified 1-methylbutyl 3-methylbutanoate as a major component of the extract from the van der Vecht gland of *V. mandarinia*. This ester elicits a strong defensive reaction in the smaller sympatric prey, *V. simillima xanthoptera*; the behaviour



Figure 1 Bioassay to evaluate alarm-pheromone activity of samples. Guard hornets rush towards filter paper impregnated with a single-worker equivalent of 2-pentanol, 3-methyl-1-butanol and 1-methylbutyl 3-methylbutanoate. The insects did not respond to filter paper impregnated only with diethyl ether.

is identical to the defensive reaction by *V. simillima* when it detects the foraging-site-marking pheromone⁴ secreted by *V. mandarinia* scouts. This 'kairomone' detection is an example of adaptation to predation by sympatric species.

The synergistic interaction and the kairomonal function described here indicate that hornets probably use semiochemicals in a complex form. Single-component alarm pheromones are found in the venom of other insects (for example, 2-methyl-3-buten-2-ol in *Vespa crabro*⁵, *N*-(3-methylbutyl)acetamide in *Vespula squamosa*⁶ and (2*S*,6*R*,8*S*)-2,8-dimethyl-1,7-dioxaspiro[5.5]undecane in *Polybia occidentalis*⁷).

A preliminary analysis of the volatile components in venom from all seven Japanese hornet species confirms that they contain alcohols and esters that are also common in manufactured food and cosmetic products (our unpublished results). It may therefore be sensible to screen these commercial products for the presence of

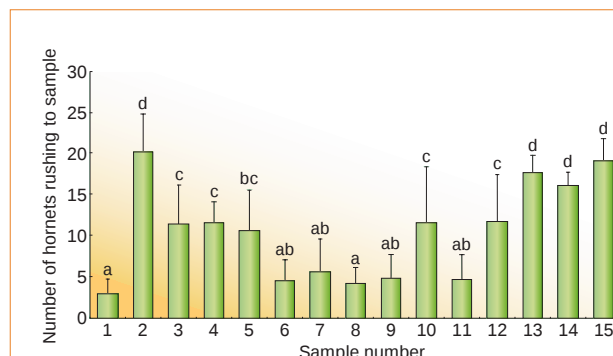


Figure 2 Response by guard hornets (*Vespa mandarinia*) to impregnated filter paper placed at the entrance to a feral giant-hornet colony containing about 200 workers. Samples 1 and 2 were 100 μ l each of diethyl ether and venom extract, respectively; samples 3–15, which were each diluted in diethyl ether (10 nl per 100 μ l) were as follows: 3, (*S*)-2-pentanol;

4, (*R*)-2-pentanol; 5, ($\pm$)-2-pentanol; 6, 3-methyl-1-butanol; 7, (*S*)-1-methylbutyl 3-methylbutanoate; 8, (*R*)-1-methylbutyl 3-methylbutanoate; 9, ($\pm$)-1-methylbutyl 3-methylbutanoate; 10, ($\pm$)-2-pentanol and ($\pm$)-1-methylbutyl 3-methylbutanoate; 11, ($\pm$)-1-methylbutyl 3-methylbutanoate and 3-methyl-1-butanol; 12, ($\pm$)-2-pentanol and 3-methyl-1-butanol; 13, (*S*)-2-pentanol, 3-methyl-1-butanol and (*S*)-1-methylbutyl 3-methylbutanoate; 14, (*R*)-2-pentanol, 3-methyl-1-butanol and (*R*)-1-methylbutyl 3-methylbutanoate; 15, ($\pm$)-2-pentanol, 3-methyl-1-butanol and ($\pm$)-1-methylbutyl 3-methylbutanoate. Venom extract and 2-pentanol each elicited a defensive reaction, which increased when the filter paper was impregnated with all three volatile components. Data are mean responses $\pm$ s.d. (five replicates); letters indicate a significant difference ($P < 0.01$) by Tukey's test.

pheromones that might act as an alarm signal to dangerous insects.

Masato Ono*, Hirokazu Terabe*, Hiroshi Hori†, Masami Sasaki*

*Centre of Excellence Integrative Human Science Program, Research Institute and Graduate School of Agriculture, and †Department of Applied Biological Chemistry, Tamagawa University, Machida, Tokyo 194-8610, Japan
e-mail: ono@agr.tamagawa.ac.jp

1. Ono, M. *Science of Japanese Hornets (Vespa spp.)* (Kaiyusha, Tokyo, 1997).
2. Shiota, H. *J. Agric. Food. Chem.* **41**, 2056–2062 (1993).
3. Rossiter, K. *Perfumer and Flavorist* **21**, 33–36 (1996).
4. Ono, M., Igarashi, T., Ohno, E. & Sasaki, M. *Nature* **377**, 334–336 (1995).
5. Veith, H. J., Koeniger, N. & Maschwitz, U. *Naturwissenschaften* **71**, 328–329 (1984).
6. Heath, R. R. & Landolt, P. J. *Experientia* **44**, 82–83 (1988).
7. Dani, F. R. et al. *Physiol. Entomol.* **25**, 363–369 (2000).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Optics

Mechanism for ‘superluminal’ tunnelling

In their discussion of a mechanism that I proposed¹ to explain the apparent superluminal (that is, faster than light) tunnelling of light pulses observed in photonic barrier experiments^{2–5}, Büttiker and Washburn⁶ used an old ‘reshaping’ argument⁷ that is at variance with my model¹ and is not supported by the bulk of the experimental tunnelling evidence. The mechanism I proposed agrees with experiment and resolves a long-standing paradox — namely, the lack of dependence of tunnelling time on barrier length for thick barriers (the Hartman effect)^{8,9}.

The reshaping mechanism that is cited

by Büttiker and Washburn⁶ requires that the tunnelled pulse be shortened and reshaped, as shown in their Fig. 1. According to this mechanism, the transmitted pulse is ‘front-loaded’ so that only the leading edge of the incident pulse survives the tunnelling event without being severely attenuated to the point that it cannot be detected⁶. A direct corollary of this description is that if the pulse has two humps, only the leading edge of the first hump is transmitted, resulting in a shortened, reshaped, single-humped pulse (Fig. 1).

In reality, most tunnelling experiments reveal transmitted pulses that are neither reshaped nor shortened^{2–5}, which is borne out by numerical simulations^{1,10}. As distortionless tunnelling is a narrow-band phenomenon, the pulses used in these experiments are of necessity long compared with the barrier width (reshaping only occurs when they are not³). Hence, the interaction is a quasi-steady-state process.

If the incident pulse is much longer than the transit time of a light front through the sample, it does not really propagate through the barrier: it simply modulates the energy stored in the barrier from the moment of turn-on. The output adiabatically follows the input with a delay proportional to the stored energy, just as a capacitor introduces delay in an electrical circuit. What is really observed then is a phase shift in an amplitude modulation.

Figure 1 shows a simulation of a tunnelling double-peaked pulse based on my model¹. There is no reshaping or pulse shortening, only delay and attenuation. The reshaping argument applies only to the very leading edge (the front, or ‘turn on’), which can propagate through the barrier before the steady-state reflectivity builds up. There are irreconcilable differences between this mechanism and the reshaping model.

The mechanism that I proposed readily

explains the Hartman effect^{8,9}. The delay of the peaks becomes independent of the barrier length because the stored energy saturates with length⁹. This delay time (which is proportional to stored energy) is not a propagation delay and should not be linked to a velocity. It tells us the time it takes for energy to be stored and released by the barrier¹⁰. Because of the analogy between the Helmholtz equation for electromagnetic waves and the Schrödinger equation for quantum particles, this mechanism also applies to quantum-mechanical tunnelling.

Herbert G. Winful

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, USA
e-mail: winful@eecs.umich.edu

1. Winful, H. G. *Phys. Rev. Lett.* **90**, 023901 (2003).
2. Steinberg, A. M., Kwiat, P. G. & Chiao, R. Y. *Phys. Rev. Lett.* **71**, 708–711 (1993).
3. Spielmann, C., Szepcok, R., Stingl, A. & Krausz, F. *Phys. Rev. Lett.* **73**, 2308–2311 (1994).
4. Nimtz, G., Enders, A. & Spieker, H. J. *Phys. I (France)* **4**, 565–570 (1994).
5. Longhi, S., Marano, M., Laporta, P. & Belmonte, M. *Phys. Rev. E* **64**, 055602 (2001).
6. Büttiker, M. & Washburn, S. *Nature* **422**, 271–272 (2003).
7. Japha, Y. & Kurizki, G. *Phys. Rev. A* **53**, 586 (1996).
8. Hartman, T. E. J. *Appl. Phys.* **33**, 3427–3433 (1962).
9. Winful, H. G. *Opt. Express* **10**, 1491–1496 (2002).
10. Winful, H. G. *J. Sel. Top. Quant. Electron.* **9**, 17–29 (2003).

Büttiker and Washburn reply — Winful objects to our appeal¹ to pulse reshaping discussed in refs 2,3. It is these discussions, however, that show that causality is maintained in the tunnelling process and that no fundamental tenet of modern physics is called into question. Although these discussions were put forth for short pulses, they are not, in principle, limited to this case. Indeed, pulse reshaping also occurs for long pulses, even if only at the leading edge of the pulse.

We contend that the positions of the peak maxima are not causal, and hence are not a meaningful quantity for the definition of speed or tunnelling time. This is true for short pulses, but even more so for longer ones: the broader a wave packet becomes, the less significant is the position of its peak. As the central point in this debate is a meaningful tunnelling time, we largely ignored peak maxima in our discussions.

Markus Büttiker*, Sean Washburn†

*Department of Theoretical Physics, University of Geneva, 234 quai Ernest Ansermet, 1211 Geneva 4, Switzerland
e-mail: buttiker@karystos.unige.ch
†Department of Physics and Astronomy, University of North Carolina, Chapel Hill, North Carolina 27599-3255, USA
e-mail: sean@physics.unc.edu

1. Büttiker, M. & Washburn, S. *Nature* **422**, 271–272 (2003).
2. Deutch, J. E. & Low, F. E. *Ann. Phys.* **288**, 184–202 (1993).
3. Japha, Y. & Kurizki, G. *Phys. Rev. A* **53**, 586–590 (1996).

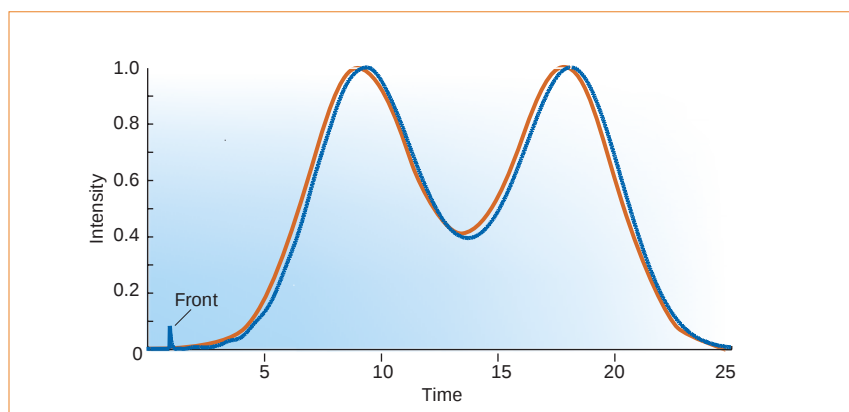


Figure 1 Tunnelling of a double-peaked pulse for a barrier consisting of a periodic dielectric structure of length L and a transmission of 1.4×10^{-3} . The delay of the tunnelled peaks (full line) relative to the input peaks (dashed line) is $0.25L/c$, or one-quarter the transit time L/c through an equivalent distance *in vacuo*. This short delay is not a propagation delay but a cavity decay time. The ‘front’ propagates at c . It has nothing to do with the actual tunnelling, which occurs far behind the light front. The mechanism cited by Büttiker and Washburn⁶ applies only to the front.

The evolution of comets in the Oort cloud and Kuiper belt

S. Alan Stern

Department of Space Studies, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, Colorado 80302, USA

Comets are remnants from the time when the outer planets formed, ~4–4.5 billion years ago. They have been in storage since then in the Oort cloud and Kuiper belt—distant regions that are so cold and sparsely populated that it was long thought that comets approaching the Sun were pristine samples from the time of Solar System formation. It is now recognized, however, that a variety of subtle but important evolutionary mechanisms operate on comets during their long storage, so they can no longer be regarded as wholly pristine.

Comets are small bodies with characteristic sizes of 1 to 15 kilometres that orbit the Sun^{1,2}. They are usually detected as they approach the Sun because their near-surface volatiles sublimate under the increasing insolation, in turn generating an extensive, highly visible gas-and-dust atmosphere, called the coma. Because of the small size of the solid nucleus of the coma, a comet's gravity is too weak to retain these constituents, so the coma expands to great distances and is lost to space. As first recognized decades ago^{3,4}, the cometary nucleus is the source of the escaping gas and dust that make up both the coma, and its extension, called the tail. Strong circumstantial evidence, based on the ease with which comets split and fragment, points to the inherent mechanical weakness of cometary nuclei⁵; in fact, many comets may essentially be strengthless, gravitationally bound 'piles of rubble'⁶.

Comets consist of approximately equal proportions of nonvolatile solids (silicates, refractory organics) and volatile ices. Cometary ices are dominated by water ice³, but CO₂ and CO are also present at significant levels (in extreme cases having combined abundances as high as ~15–20% that of the water ice²). Other volatiles, notably including H₂S, CH₃OH, H₂CO, NH₃, HCN, CH₄ and S₂, have also been detected in the atmospheres of comets. The presence of such a wide array of high-volatility species strongly suggest that comets (1) originated in the cool, outer regions of the Sun's protoplanetary nebula and (2) have ever since been stored only in cold conditions^{2–4}.

Most comets have very long-period orbits that extend from thousands to several tens of thousands of astronomical units (AU) from the Sun. This, and the related observation that the orbits of such comets are nearly isotropically oriented relative to the plane of

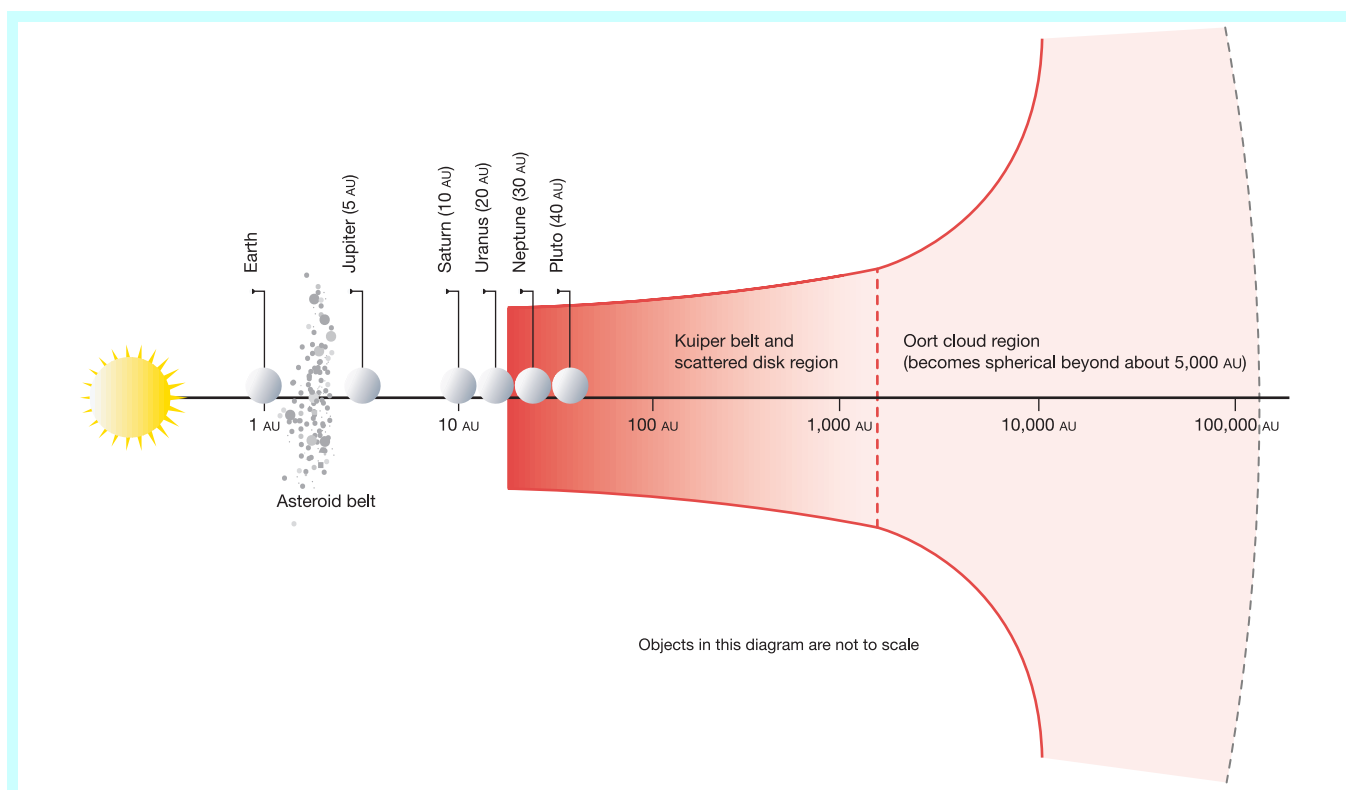


Figure 1 Diagram showing the Kuiper belt and Oort cloud to scale with our planetary system. The huge scale of the Oort cloud can be appreciated by the fact that the

nearest star is located only about three times farther from the Sun than the edge of the Oort cloud (that is, $\sim 3 \times 10^5$ AU from the Sun).

the Solar System, caused Oort to conclude in 1950 that such comets must be derived from an essentially spherical reservoir surrounding the Sun at these very great distances⁸. This reservoir is now known as the Oort cloud. Modern estimates⁹ place the number of Oort cloud comets in the range 10^{11} to perhaps 5×10^{12} , corresponding to a total Oort cloud mass of order $1 M_{\oplus}$ (where $M_{\oplus}$ is the mass of the Earth) to perhaps $50 M_{\oplus}$ (depending also upon the presently ill-determined typical masses of cometary nuclei). Models of Solar System and Oort cloud formation^{10,11} have repeatedly shown that the formation of an Oort cloud is a natural by-product of the clearing and ejection of debris from the giant planets' region some ~ 3.5 – 4.5 Gyr ago. (See Table 1.)

Not all comets, however, have the highly extended orbits indicative of an extremely distant, spheroidal reservoir. On the contrary, many comets are observed to be on much more tightly bound orbits that are either trapped among the planets, or never stray to the Oort cloud¹². Unlike the isotropically distributed orbits of the Oort cloud comets, these shorter-period comets predominantly display shallow, prograde orbital inclinations relative to the plane of the Solar System.

In the late 1980s and early 1990s, these facts were used to infer that a second, as yet undiscovered cometary reservoir must exist. This second reservoir must be far more compact (lying primarily within a few hundred AU) and more disk-like (typical orbital inclinations of $<30^\circ$) than the huge, spherical Oort cloud¹³. The existence of such a disk of material beyond the planets had been suggested by various workers, most famously Kuiper¹⁴. Kuiper's hypothesis was that a debris field of remnant, small bodies might naturally be expected beyond the last of the giant planets, representing a region where planetary formation had not gone to completion. By the early 1990s it became possible for groundbased techniques to detect moderate-sized bodies in this region directly, which is now known as the Kuiper belt¹⁵.

Comets are the oldest and coldest, the most accessible, and the most nearly pristine samples of the outer solar nebula available for study². Of course, it was known even 20 years ago that comets are modified by early, internal radiogenic heating⁷ and insolation-driven evolution after dislodgement from their storage regions¹⁶. It was long thought that comets were essentially held in perfect stasis during their long storage in the remote, cold dynamical storage reservoirs known as the Oort cloud and the Kuiper belt. In recent years, however, a broad array of processes have been discovered that could describe the evolution of comets while they are in cold storage. Owing to our lack of knowledge of the thermophysical and structural details of cometary surfaces, which can only be definitively explored through *in situ* orbiter/lander and sample return missions, it is not yet clear which of these evolutionary processes dominate. Nonetheless, there is a growing appreciation that comets do evolve during storage in the Oort cloud and the Kuiper belt in a variety of potentially important ways.

Processes affecting comets during dynamical storage

The longstanding big-picture view of cometary evolution has been that the cryogenic conditions and highly dilute nature of the Oort cloud and the Kuiper belt suggest conditions of stasis. This, along with evidence that comets themselves are samples of the planetesi-

mal population from which the planets were built, suggested long ago¹⁷ that comets represent wholly pristine samples from the formation epoch of the planets. Chiefly for these reasons, comets remain on the top priority list of planetary science mission targets¹⁸. Nonetheless, a variety of thermal, collisional, radiation, and interstellar medium (ISM) processes affect comets during their long storage in the Oort cloud and Kuiper belt. These processes complicate the interpretation of cometary observations and suggest that *in situ* sample analysis and sample return missions will reveal a substantially modified surface layer on comets. I will now describe each of the various effects that cause comets to evolve while in dynamical storage far from the Sun.

Thermal processes. Because of the Oort cloud's immense cross-section, it has been recognized since the 1950s that passing stars regularly penetrate it; this process has long been known to be fundamental to the diffusion and dynamical randomization of orbital inclinations in the Cloud^{8,11,19}. In the late 1980s, it was also recognized that such encounters also have consequences for the heating of comets there²⁰.

Importantly, rare parsec-range and closer encounters with highly luminous O and supergiant stars ($L_{\star} \approx 3$ – $6 \times 10^5 L_{\odot}$, where L is the luminosity of a star or the Sun, depending on the subscript) as far away as 5 pc were found to be capable of heating the entire Oort cloud well above its ambient 5–6 K temperature, to temperatures capable of removing the most volatile ices from surface layers. For example, there is a unit chance that in the past 4.5 Gyr a 'nearby' (5 pc distant) O star has heated the Oort cloud to 16 K, thereby removing species like condensed neon and molecular oxygen. (Owing to the far greater proximity of the Sun to the Kuiper belt than the Oort cloud, however, such highly luminous stars do not much affect the 30–60 K ambient temperature of Kuiper belt comets, raising them only ~ 1 K.)

Although supernovae explosions are far briefer (~ 0.1 yr) than O-star passages ($\sim 3 \times 10^4$ yr), supernovae are so much more luminous ($L_{\star} \approx 10^9 L_{\odot}$) than even the brightest O stars, that they can heat the Oort cloud from far larger distances. Using modern supernova rates and luminosity estimates, it has been estimated²⁰ that ~ 30 supernovae heating events close enough (<20 pc) to heat the surfaces of Oort cloud objects to 30 K have occurred in the past 4 Gyr; this in turn depletes the abundances of condensed, near-surface argon, CO, N₂ and CH₄. Furthermore, models indicate there is a unit chance that the Oort cloud should have experienced one supernova heating event to 50 K over the past 4 Gyr, and there is a 50% probability that a supernova heating event occurred to warm all cometary surfaces to 60 K, at which temperature other important species like formaldehyde will be depleted. This effect probably masks some primordial signatures in the surface (but not the deep interior) compositions of Oort cloud and Kuiper belt comets.

Given a heating timescale, the depth of penetration of the thermal wave resulting from a stellar or supernova encounter can be estimated from the diffusion equation, based on the thermal properties of the surface material, the orientation of the cometary rotation pole, and the duration of the heat pulse. The major uncertainty in this calculation is the thermal diffusivity of cometary

Table 1 The primary cometary reservoirs of the Solar System

	Kuiper belt	Oort cloud
Shape	Disk-like	Spheroidal
Distance range	30–1,000 AU	1×10^3 – 1×10^5 AU
Comet population	~ 5 – 10×10^9	1×10^{11} – 5×10^{12}
Estimated mass (including smaller debris)	$\sim 0.1 M_{\oplus}$	1 – $50 M_{\oplus}$
Ambient surface temperatures	30–60 K	5–6 K
Origin	Largely <i>in situ</i>	Ejected material from the Kuiper belt and outer-planets zone
Return mechanism from the reservoir	Dynamical chaos due to planetary perturbations and collisions	Perturbations due to passing stars, galactic tides and molecular clouds

surfaces, which—owing to the lack of returned comet samples and orbital/lander missions—has never been directly measured. By adopting a range of surface diffusivities consistent with ice-conductivity measurements relevant to cometary materials^{21,22}, it was found that the thermal wave of a typical, O-star encounter could plausibly penetrate to depths of 5–50 m into comets. Supernovae encounters, while generating a more severe thermal pulse, are much briefer. Therefore their intense but shorter thermal pulses have been estimated to propagate only 0.1–2 m into cometary surfaces²⁰. The primary effect of these heating events is preferentially to remove supervolatiles like O₂, N₂, He, Ne, CO, CH₄ and Ar from cometary surface layers. However, these heating events may also be related to both high cometary ice spin-temperature results^{2,23} and the depletion of argon recently reported in some comets by observers using the FUSE spacecraft²⁴.

Collisional processes. The enormous volume of the Oort cloud dilutes the spatial density of comets to very low levels (the mean separation between comets in the Oort cloud is of the order of 50–500 million km). This, combined with the very low orbital velocities in the Oort cloud (of the order of 0.2 km s^{−1}), naturally suggests an almost collisionless environment. Collision-rate estimates for bodies in the Oort cloud²⁵ have confirmed this, indicating that (1) Oort cloud comets must be ancient and (2) the fraction of a typical Oort cloud comet's surface that is expected to be covered by craters caused by collisions with other objects during the 4.5-Gyr storage in the Oort cloud is of order 1% or less, depending on assumptions regarding the Cloud's spatial density. Nonetheless, although collisions in the Oort cloud itself are rare and apparently of little effect, recent work indicates that collisions probably greatly affected the surfaces and interiors of many of these bodies during their ejection from the planetary region to the Oort cloud²⁶.

The collisional history of the Kuiper belt is different^{27,28}. Kuiper belt comets are widely thought to have originated essentially *in situ* in the Kuiper belt. However, owing to the higher orbital speeds (typically 4 km s^{−1}) and space density of kilometre-sized bodies (higher than in the Oort cloud by a factor of 2 × 10⁵), collision rates in the KB exceed those in the Oort cloud by a factor of the order of 10⁶. Importantly, Kuiper belt collisions happen at such high speeds that they are highly erosional, causing Kuiper belt comets to lose surface material over time. In fact, the time-averaged rate of erosion in the Kuiper belt is so high that almost all kilometre-scale and smaller Kuiper belt bodies are now thought to be fragments 'chipped' off larger Kuiper belt objects in the last 10–20% of the age of the Solar System. Hence, most comets derived from the Kuiper belt are expected to exhibit low surface exposure ages, and are therefore not expected to exhibit many craters.

These results, now obtained by many independent groups^{27,28}, represent both a significant shift in our view of cometary bodies, and a great difference between Kuiper belt and Oort cloud comets. Whereas Oort cloud comets are still expected to be (damaged) relics

of the formation era, most Kuiper belt comets must be young. Furthermore, owing to collisions, the interior mechanical properties and structure of both Oort cloud and Kuiper belt comets are unlikely to reflect their gentle accretional environment, having been significantly modified by the effects of collisions thereafter.

Interactions with the interstellar medium. The first study of the interaction between comets and the ISM was made three decades ago^{29,30}, when it was suggested that ISM gas would slowly accrete onto cometary surfaces, forming an accretion crust that might amount to a layer 10–100 μm thick over 4.5 Gyr. However, these studies ignored the role of high-velocity ISM grain impacts on cometary surfaces.

Later, when the competition between ISM grain-driven erosion and ISM gas-driven accretion was first modelled³¹, it was found that grain-driven erosion on icy surfaces is about 1,000 times more efficient than gas accretion, causing comets to lose (rather than gain) material as a result of ISM interactions. The primary reason for this is that while gas sticking is inefficient (typical sticking ratios being a few per cent), high-velocity grain impacts (at typical ISM–comet relative velocities of 10–30 km s^{−1}) are very efficient at micro-cratering (causing far more material to be lost from the cometary surface than the imported mass of the impacting grain). Similar results were obtained in studying the source of dust generated in β Pictoris-type stellar disks³².

Owing primarily to the strong density gradients between the cloud and intercloud phases of the interstellar medium, it was subsequently found³³ that ISM interactions vary severely in time. Indeed, it was found that the majority of the surface erosion comes during the occasional passage of the Solar System through galactic, giant molecular clouds (GMCs). It is estimated that about 5–10 such erosion events have occurred since the formation of the Oort cloud, which may in total have caused comets in the Oort cloud to have lost about 1–20 m of surface material since the formation of the Oort cloud².

Comets in the Kuiper belt will also suffer from this erosion because they cannot be protected by the heliosphere (ISM pressure during GMC passage compresses the heliosphere down to a size perhaps as small as 5 AU). However, given the much younger ages of cometary surfaces in the Kuiper belt (owing to collisions), the number of expected GMC erosion events likely to have occurred on these bodies is only about one, and could plausibly be zero. Therefore, it is uncertain whether the surfaces of present-day Kuiper belt comets may, or may not, have experienced a significant ISM-driven surface erosion.

Radiation processes. This area subdivides naturally into two sub-categories: photon bombardment and charged particle bombardment³⁴. Considering photon bombardment first, interstellar and solar ultraviolet (that is, $h\nu > 3$ eV) photons have copious fluxes in the Oort cloud and Kuiper belt, providing the energy necessary to break bonds and initiate substantial chemical change in cometary

Table 2 Cometary evolution mechanisms in the Kuiper belt and Oort cloud

	Primary effects	Maximum modification depth	Primary temporal style	Notes
Heating by supernovae	Loss of supervolatiles	Metres	Stochastic	Less important in the Kuiper belt
Heating by passing stars	Loss of supervolatiles	Tens of metres	Stochastic	Less important in the Kuiper belt
Collisions	Cratering, regolith evolution and overturning, and structural/thermophysical evolution	Tens of metres	Stochastic	Not important in the Oort cloud
Ultraviolet damage	Chemical reactions, polymerization, surface albedo and colour changes	Tens of micrometres	Continuous	More effective in the Kuiper belt
Cosmic ray damage	Chemical reactions, sputtering devolatilization, polymerization, surface albedo, colour and microstructure changes	Metres	Continuous	—
Sputtering erosion by ISM gas	Regolith removal and selective loss of volatiles	Tens of micrometres	Continuous	More effective in the Kuiper belt
Mechanical erosion by ISM grains	Regolith removal and erasure of other evolutionary effects	Metres	Stochastic	—

surfaces. Ultraviolet photospattering is capable of eroding away the uppermost few micrometres of icy surfaces³⁵. But more importantly, in a classic series of laboratory experiments and theoretical studies, M. Greenberg showed that ultraviolet photons would produce significant alteration of the composition, colour, and volatility of the upper several to few tens of micrometres of cometary surfaces³⁶. Others^{37,38} confirmed and extended these results, showing that ultraviolet photons promote surface darkening (to albedos of only a few per cent) and devolatilization that becomes progressively more severe with dosage, and therefore age. Because of their much closer proximity to the Sun, Kuiper belt comets experience a much ($\sim 10^5$ times) higher ultraviolet and solar cosmic ray (SCR) surface dose, greatly increasing the total deposited charged-particle energy incident on the surfaces of these bodies, relative to Oort cloud comets, but their ~ 10 times lower average surface age somewhat mitigates this effect.

Now consider energetic charged particles. The fluence of charged particles in the Oort cloud is dominated by Galactic Cosmic Rays (GCRs) with keV-to-MeV energies. Like ultraviolet photons, charged particle radiation is capable of both sputtering surfaces and breaking bonds, thereby inducing chemical reactions and consequently reordering the surface ice matrix³⁹. The irreversible radiation-driven conversion of the original, water-dominated ice matrix to a more complex 'crust' inevitably leads to the darkening of the surface, via the formation of long-chain hydrocarbons^{36,37}. If cometary surfaces have bulk densities of 1 g cm^{-3} or less, then the cosmic ray damage layer may reach several metres in depth⁴⁰. The GCR dose onto Kuiper belt comets is expected to be much reduced, relative to Oort cloud comets, because of the shielding effects of the heliosphere out to $\sim 100 \text{ AU}$.

The chemical and structural changes in cometary surfaces produced by radiation processes may be responsible for some instances of cometary activity at large distances (5–15 AU) during the approach of new comets to perihelion¹⁶. Indeed, this may explain the greater tendency for Oort cloud comets to erupt at large heliocentric distances on their first approach to the Sun¹⁶. Interestingly, sputtering by cosmic rays, solar wind, or perhaps even hot ISM gas may also be responsible for generating the particles that, after acceleration, explain the long-mysterious anomalous cosmic rays (ACRs)⁴¹.

Towards a new perspective

As a result of the discovery of evolutionary processes acting in the Oort cloud and the Kuiper belt, comets—though still the most pristine bodies known—have been modified in several important ways since their birth. It also now seems inevitable that most comets from the Kuiper belt, although they are constructed of ancient material, cannot themselves be ancient—instead they must be 'recently' created chips off larger Kuiper belt objects, formed in violent (km s^{-1} class) impacts, rather than the gentle environment required for cometary accretion in the early solar nebula. (See Table 2.)

Beyond deepening our knowledge of comets and their storage reservoirs, these effects also indicate a fascinating link between long-term cometary surface evolution and the Sun's 4.6-Gyr passage through the ISM and the Galaxy. Moreover, it now appears that the surfaces, surface ages, and deep interior mechanical properties of comets derived from the Kuiper belt and Oort cloud could be different from one another, at least before their modification by intense activity associated with heating due to passages into the warm inner Solar System.

The realization that comets do slowly evolve during their long storage in remote dynamical reservoirs provides insight and context to more confidently evaluate the results of astronomical and space mission observations of comets. It also strongly argues for relatively deep (metres scale or deeper) subsurface sampling of comets, if pristine samples of ancient material are to someday be had. □

doi:10.1038/nature01725.

- Whipple, F. L. Origin of the comets. *Proc. Natl Acad. Sci. USA* **51**, 711–715 (1964).
- Mumma, M. L., Weissman, P. R. & Stern, S. A. in *Protostars and Planets III* (eds Levy, E. H., Lunine, J. I. & Matthews, M. S.) 1177–1252 (Univ. Arizona Press, Tucson, 1993).
- Whipple, F. L. A comet model I: The acceleration of comet Encke. *Astrophys. J.* **111**, 375–394 (1950).
- Delsemme, A. H. in *Comets, Asteroids, and Meteors: Interrelations, Evolution, and Origins* (ed. Delsemme, A. H.) 3–12 (Univ. Toledo, Toledo, 1977).
- Asphaug, E. & Benz, W. Size, density, and structure of comet Shoemaker-Levy 9 inferred from the physics of tidal breakup. *Icarus* **121**, 225–248 (1996).
- Weissman, P. R., Asphaug, E. & Lowry, S. C. in *Comets II* (eds Festou, M. C., Keller, U. & Weaver, H. A.) (Univ. Arizona Press, Tucson, in the press).
- Prialnik, D. & Podolak, M. Changes in the structure of comet nuclei due to radioactive heating. *Space Sci. Rev.* **90**, 169–215 (1999).
- Oort, J. H. The structure of a cloud of comets surrounding the solar system and a hypothesis concerning its structure. *Bull. Astron. Inst. Neth.* **11**, 91–110 (1950).
- Weissman, P. R. in *Completing the Inventory of the Solar System* (eds Rettig, T. W. & Hahn, J. M.) *ASP Conf. Proc.* **107**, 265–288 (1996).
- Safronov, V. S. in *Evolution of the Small Bodies of the Solar System* (eds Fugichioni, M. & Kresak, K.) 217–226 (North-Holland, Amsterdam, 1987).
- Dones, L. et al. Simulations of the formation of the Oort Cloud. *Icarus* (submitted).
- Levison, H. F. & Duncan, M. J. The long-term dynamical behavior of short-period comets. *Icarus* **108**, 18–36 (1994).
- Duncan, M. J., Quinn, T. & Tremaine, S. The origin of the short period comets. *Astrophys. J.* **328**, L69–L73 (1988).
- Kuiper, G. P. in *Astrophysics: A Topical Symposium* (ed. Hynek, J. A.) 357–424 (McGraw Hill, New York, 1951).
- Jewitt, D. C. & Luu, J. X. The discovery of the candidate Kuiper Belt object 1992QB₁. *Nature* **362**, 730–732 (1993).
- Meech, K. J. Chemical and physical aging of comets. *IAU Colloq* **173**, (eds Svoren, J., Pittich, E. M. & Rickman, H.) 195–201 (Astronomical Institute of the Slovak Academy of Sciences, Tatranska Lomnica, 1999).
- Delsemme, A. H. in *Comets* (ed. Wilkening, L. L.) 85–131 (Univ. Arizona Press, Tucson, 1982).
- Belton, M. et al. *New Frontiers in the Solar System: An Integrated Exploration Strategy. Solar System Exploration Survey 457* (National Academies Press, Washington, 2002).
- Weissman, P. R. The Oort Cloud. *Nature* **344**, 825–830 (1990).
- Stern, S. A. & Shull, J. M. The influence of supernovae and passing stars on comets in the Oort cloud. *Nature* **332**, 407–411 (1988).
- Klinger, J. in *Ices in the Solar System* (eds Klinger, J., Benest, D., Dolfus, A. & Smoluchowski, R.) 407–417 (D. Reidel, Dordrecht, 1985).
- Kouchi, A. et al. Extremely low thermal conductivity of amorphous ice: Relevance to comet evolution. *Astrophys. J.* **388**, L73–L76 (1992).
- Mumma, M. J. in *From Stardust to Planetesimals* (eds Pendleton, Y. J. & Tielens, A. G. G. M.) *ASP Conf. Ser.* **122**, 369–396 (1997).
- Weaver, H. A. et al. A Search for argon and O VI in three comets using the Far Ultraviolet Spectroscopic Explorer. *Astrophys. J.* **576**, L95–L98 (2002).
- Stern, S. A. Collisions in the Oort Cloud. *Icarus* **73**, 499–507 (1988).
- Stern, S. A. & Weissman, P. R. Rapid collisional evolution of comets during the formation of the Oort cloud. *Nature* **401**, 589–591 (2001).
- Farinella, P., Davis, D. R. & Stern, S. A. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 125–133 (Univ. Arizona Press, Tucson, 2000).
- Durda, D. D. & Stern, S. A. Collision rates in the present-day Kuiper Belt and Centaur Regions: Applications to surface activation and modification on comets, KBOs, and Pluto-Charon. *Icarus* **45**, 220–229 (2000).
- O'Del, C. R. A new model for cometary nuclei. *Icarus* **19**, 137–146 (1973).
- Whipple, F. L. in *Comets, Asteroids and Meteors* (ed. Delsemme, A.) 25–34 (Univ. Toledo Press, Toledo, 1977).
- Stern, S. A. The effects of mechanical interaction between the interstellar medium and comets. *Icarus* **68**, 276–283 (1986).
- Lissauer, J. J. & Griffith, C. A. Erosion of circumstellar particle disks by interstellar dust. *Astrophys. J.* **340**, 468–471 (1989).
- Stern, S. A. ISM-induced erosion and gas-dynamical drag in the Oort Cloud. *Icarus* **84**, 447–466 (1990).
- Johnson, R. E. *Energetic Charged-Particle Interactions with Atmospheres and Surfaces* 232 (Springer, New York, 1990).
- Throop, H. B. *Light Scattering and Evolution of Protoplanetary Disks and Planetary Rings* 1–162, (PhD thesis, Univ. Colorado (2000)).
- Greenberg, J. M. in *Comets* (ed. Wilkening, L. L.) 131–164 (Univ. Arizona Press, Tucson, 1982).
- Thompson, W. R., Murray, B. G. J. P. T., Khare, B. N. & Sagan, C. Coloration and darkening of methane clathrate and other ices by charged particle irradiation: Applications to the outer solar system. *J. Geophys. Res.* **92**, 14933–14947 (1987).
- Hudson, R. L. & Moore, M. H. Energetic processing of laboratory ice analogs: UV photolysis versus ion bombardment. *J. Geophys. Res.* **106**, 33381–33386 (2001).
- Gerakines, P. A., Moore, M. H. & Hudson, R. L. Radiation chemical alterations in solar system ices: An overview. *J. Geophys. Res.* **106**, 33275–33284 (2001).
- Strazzula, G. in *Composition and Origin of Cometary Materials* (eds Altwegg, K., Ehrenfreund, P., Geiss, J. & Huebner, W.) 269–274 (Kluwer, Dordrecht, 1999).
- Schwandron, N. et al. The Outer source of pickup ions and anomalous cosmic rays. *Geophys. Res. Lett.* **29**, 1993–1996 (2002).

Acknowledgements I thank D. Boice, L. Dones, D. Durda, K. Meech, H. Throop and P. Weissman for comments on this article.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to S.A.S. (astern@swri.edu).

Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy

Koji Yonekura^{1,2,3*}, Saori Maki-Yonekura^{1,3*} & Keiichi Namba^{1,2,3}

¹Protonic NanoMachine Project, ERATO, JST,

²Graduate School of Frontier Biosciences, Osaka University, and

³Dynamic NanoMachine Project, ICORP, JST, 3-4 Hikaridai, Seika, Kyoto 619-0237, Japan

* These authors contributed equally to this work

The bacterial flagellar filament is a helical propeller for bacterial locomotion. It is a helical assembly of a single protein, flagellin, and its tubular structure is formed by 11 protofilaments in two distinct conformations, L- and R-type, for supercoiling. The X-ray crystal structure of a flagellin fragment lacking about 100 terminal residues revealed the protofilament structure, but the full filament structure is still essential for understanding the mechanism of supercoiling and polymerization. Here we report a complete atomic model of the R-type filament by electron cryomicroscopy. A density map obtained from image data up to 4 Å resolution shows the feature of α -helical backbone and some large side chains. The atomic model built on the map reveals intricate molecular packing and an α -helical coiled coil formed by the terminal chains in the inner core of the filament, with its intersubunit hydrophobic interactions having an important role in stabilizing the filament.

Bacteria swim by rotating helical flagellar filaments, which grow as long as 15 μm , but the diameter is only 120–250 Å. The rotary motor at the base of the filament drives the rotation of this helical propeller^{1,2} at hundreds of revolutions per second^{3,4}. For chemotaxis and thermotaxis, the swimming pattern of bacteria such as *Salmonella* and *Escherichia coli* alternates between ‘run’ and ‘tumble’; a run lasts for a few seconds and a tumble for a fraction of second. During a run, the motor rotates anticlockwise (as it is viewed from outside the cell), and several flagellar filaments with a left-handed helical shape form a bundle and propel the cell. A tumble is caused by quick reversal of the motor to clockwise rotation⁵, which produces a twisting force that transforms the left-handed helical form of the filament into a right-handed one^{6,7}, causing the bundle to fall apart rapidly. The separated filaments act in an uncoordinated way to generate forces that change the orientation of the cell. Thus, the structure of the flagellar filament and its dynamic properties have an essential role in bacterial taxis.

The filament is a helical assembly of flagellin with roughly 11 subunits per two turns of the 1-start helix; it can also be described as a tubular structure comprising 11 protofilaments, which are nearly longitudinal helical arrays of subunits⁸. Left- and right-handed helical forms are produced by supercoiling caused by a mixture of two distinct protofilament conformations, L- and R-type^{9–12}. When all 11 protofilaments are of the same type, two types of straight filaments with distinct helical symmetries are formed: the longitudinal 11-start helix is left-handed in the L-type and right-handed in the R-type straight filament.

Electron cryomicroscopy and X-ray fibre diffraction have revealed the domain organization of flagellin and subunit packing in these two straight filaments at about 10 Å resolution^{13–18}; however, higher resolution is needed to understand the structural basis of the filament formation and supercoiling in atomic detail. Flagellin has a strong tendency to polymerize into filaments and this has prevented its crystallization. By clipping off terminal chains of about 100 residues in total, fragment F41 has been crystallized and the structure solved at 2.0 Å resolution by X-ray crystallography¹⁹. The crystal structure revealed the R-type protofilament

structure consisting of the F41 subunit, which is composed of three domains, D1, D2 and D3. Domain D1 forms the outer core of the filament and domains D2 and D3 form the projection on the filament surface. Simulated extension of the R-type protofilament model showed a small but significant conformational change of the β -hairpin in domain D1, which covers most of the axial molecular interface in the F41 protofilament, and it was interpreted as the structural mechanism of switching from the R- to L-type protofilament conformation. Atomic models of the two straight filaments have been built with the atomic model of F41 by combining all available data from X-ray crystallography, X-ray fibre diffraction, and electron cryomicroscopy, and an insight into the twist-curvature coupling for the supercoiling was obtained from lateral interactions between domains D1 of neighbouring protofilaments (K. Imada *et al.*, manuscript in preparation). However, the terminal chains in the inner core, which have important roles in filament formation and assembly regulation, are still missing in these models. Thus the complete atomic model of the filament is still essential for understanding the mechanism of supercoiling and filament formation in detail.

We carried out structure analysis of the R-type straight filament by electron cryomicroscopy and helical image reconstruction. An obtained density map allowed us to locate and orient the three domains of F41 accurately and to trace the terminal chains in the inner core. Here we present a complete atomic model of the R-type filament, as the first atomic model of a protein molecule obtained by electron cryomicroscopy and image analysis alone.

Structure determination

We used the R-type straight filament reconstituted from flagellin with a point mutation of Ala 449 to Val^{20,21}. We only used images of frozen-hydrated filaments showing strong and sharp layer-lines in the Fourier transform. Distortion correction²² and solvent flattening²³ were applied to individual images in a similar way described previously, with some significant improvements and modifications in the procedure. The total number of the filament images used for the reconstruction was 102, and the number of molecular images

averaged was approximately 41,000. For the three-dimensional image reconstruction, we included layer-line data within an ellipsoid in the Fourier space that covers a resolution of 4.0 Å in the direction of the filament axis and 5.0 Å in the equatorial direction. The map was calculated by helical image reconstruction with the parameters listed in Supplementary Table 1. The density map in Fig. 1a demonstrates that α -helices and β -sheets are clearly resolved throughout the molecule, although individual strands of β -sheets in domains D2 and D3 are rather difficult to identify without superimposing the atomic model of F41. Especially in the filament core domains, D0 and D1, the map revealed the path of α -helical backbone and even some large side chains (Fig. 1b and Supplementary Fig. 1).

Flagellin isolated from the SJW1103 strain of *Salmonella typhimurium* consists of 494 amino acid residues. The atomic model of the F41 fragment¹⁹ includes 395 residues from Asn 56 to Arg 450 and lacks 55 amino-terminal and 44 carboxy-terminal residues. We first docked the F41 model with its three domains, D1, D2 and D3, into the density map as a rigid body, but we found that small modifications were necessary. We moved and reoriented domain D3 to fit into the map and also modified the conformation of the two ends of α -helices in the terminal regions of F41, which are both in domain D1, extended these α -helices further down, and then traced the missing terminal chains with large side chains as fiducial in the spoke region and domain D0. The positions and orientations of densities assigned to several large side chains allowed unambiguous model building of the terminal chains. The complete atomic model thus built was refined by the program FEX-PLOR, which is an extension of FX-PLOR²⁴. FX-PLOR is a version of X-PLOR that was modified to refine the structure of macromolecular assemblies with helical symmetry. We implemented amplitude-weighted phase

residuals so that an atomic model can be refined against phases instead of amplitudes. To check the validity of the atomic model against the density map, figure of merit (FOM) obtained by combination of the electron microscopy phase and the model phase was calculated as defined in equation (1) of Supplementary Table 2.

Structure of flagellin and the filament

Flagellin consists of four linearly connected domains labelled D0, D1, D2 and D3, which are arranged from the inside to outside of the filament (Fig. 1a). The N-terminal chain starts from D0, going through D1, D2 and reaches D3, and then comes back through D2 and D1, and the C-terminal chain ends in D0. Although all three domain connections are formed by pairs of short antiparallel chains, the one that connects domains D0 and D1 is longer than the other two, and therefore it is called the spoke region. Flagellin in Fig. 1a is viewed perpendicular to the filament axis. The overall shape of flagellin looks like an upper case Greek gamma (Γ) with a vertical dimension of about 140 Å and a horizontal dimension of about 110 Å. Throughout this paper, whenever the structure is viewed in this direction, the top and the bottom corresponds to the distal and proximal side of the flagellum, respectively. We also define the distal side as 'up' and proximal side as 'down' and describe the structure accordingly.

The ribbon diagram of the C α backbone in Fig. 2a shows the chain folding. The terminal chains form an α -helical coiled coil in domain D0. The N-terminal α -helix (labelled ND0 in Fig. 2a) starts from Gln 2 and extends up to Ser 32, whereas the C-terminal α -helix (CD0) starts from Ala 459 and extends down to Ser 491. The spoke region consists of two chains (NS and CS), one from Ser 32 to Ala 44 and the other from Glu 454 to Ala 459. The N-terminal α -helix in

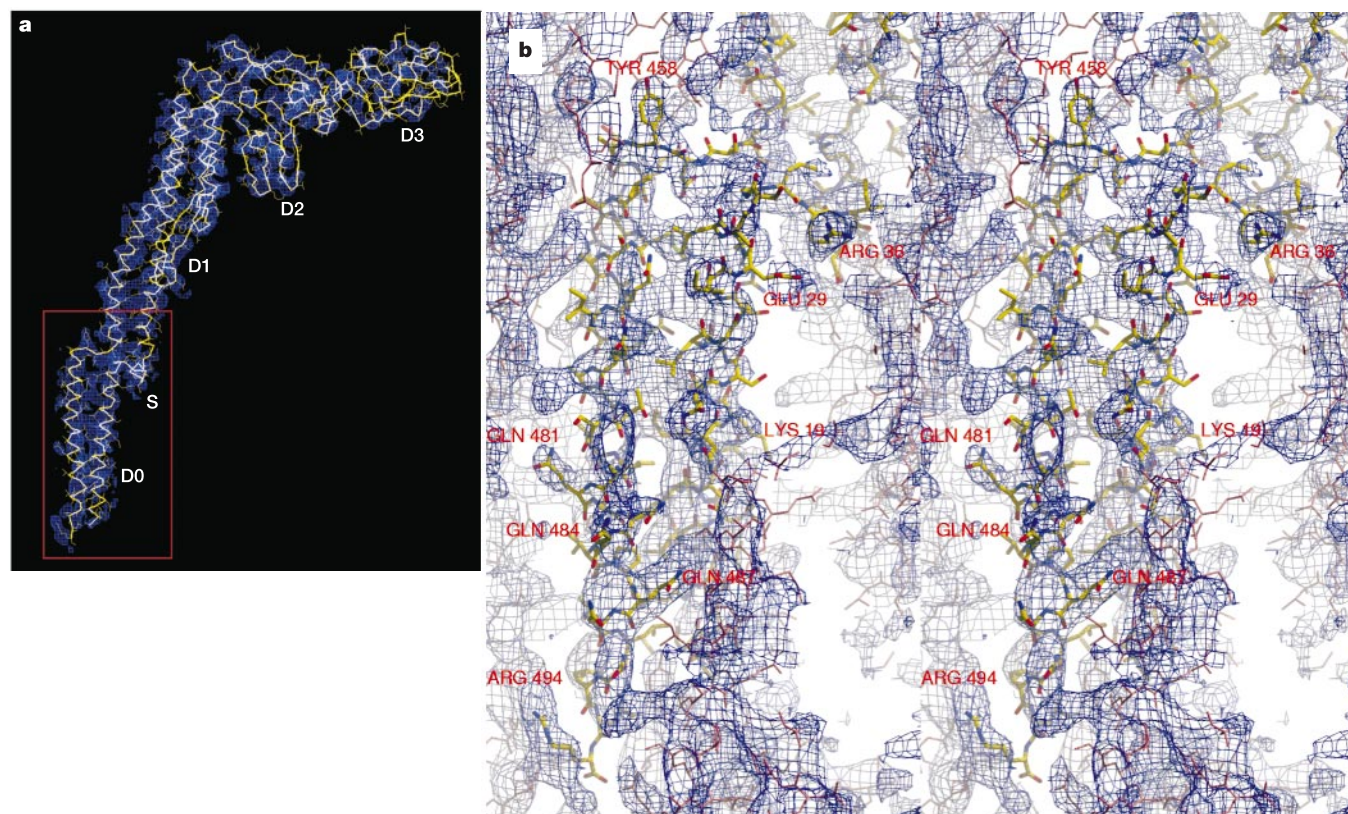


Figure 1 Density maps of the flagellar filament with the atomic model of full-length flagellin superimposed. **a**, The whole molecule, prepared with O⁴¹. Domains D0, D1, D2, D3 and the spoke region are labelled D0, D1, D2, D3 and S, respectively. **b**, Magnified stereo view of the terminal chains in domain D0. The magnified area is indicated by a box

in **a**. The atomic model of a subunit at the centre is drawn in stick representation and atoms are coloured as: carbon, yellow; nitrogen, blue; oxygen, red. The other subunits surrounding it are drawn in thin wire representation and are coloured pink. The figure was prepared with XTALVIEW⁴⁵ and RASTER3D⁴⁶. Contour levels of the maps are about 2 σ .

domain D1 (ND1a) starts from Ala 44, which is 13 residues earlier than that of F41, and extends up to Ala 99. This is followed by a loop connecting to the second, shorter α -helix (ND1b), which goes down, and the chain continues to two β -turns, a β -hairpin pointing down and an extended chain going up, and then the chain finally goes into domain D2. After ND1a, the rest of the chain folding in D1, D2 and D3 is the same as described in the F41 crystal structure¹⁹. The C-terminal α -helix in domain D1 (CD1) starts from Asn 406 and extends down to Glu 454, being longer than that of F41 by seven residues. Compared with F41, the extension of ND1a and CD1 α -helices at the bottom of domain D1 makes this portion a two-stranded α -helical coiled coil, extending the hydrophobic core of domain D1 further down (Fig. 2b). The vertical dimensions of domain D0 and D1 are about 50 Å and 80 Å, respectively. The extended chains in the spoke region are about 20 Å long.

The end-on view of the filament from the distal end of the flagellum (Fig. 3a) shows clearly the concentric double-tubular structure made of domains D0 and D1 in the densely packed filament core^{13–15}. Domains D2 and D3, which project out from the filament core, are relatively well separated from one another. The diameter of the filament is approximately 240 Å and that of the central channel is about 20 Å. The N- and C-terminal α -helices (ND0 and CD0) are radially arranged with ND0 outside and CD0 exposed to the central channel. The chains connecting domains D0 and D1 (NS and CS) look exactly like radial spokes in this view. As shown in side views from outside and inside the filament (Fig. 3b, c),

domains D0 and D1 make intimate intersubunit interactions, both axially and laterally. As a polar assembly of flagellin having a largely asymmetric structure, the filament model shows the concaved feature at the distal end (top of Fig. 3c) and the pointed tip at the proximal end (bottom), which are observed in the electron micrographs of the negatively stained filaments²⁵. The terminal chains of about 65 N-terminal and 45 C-terminal residues are unfolded in the monomeric form of flagellin in solution²⁶. It can be inferred from this structure that, in the absence of axial and lateral packing interactions of these terminal chains in the inner tube, the two-stranded α -helical coiled coil in domain D0—with relatively less extensive hydrophobic core and a pair of extended, rather flexible looking chains connecting to domain D1—would be highly unstable. In contrast, the upper two-thirds of domain D1 can form its compact tertiary structure in the monomeric form because of its extensive hydrophobic core formed by three α -helices and one β -hairpin.

Comparison with the F41 structure

Direct comparison between the crystal structure of F41 placed in the filament density map in the initial step of the model building and the structure of flagellin in the final filament model is shown in Fig. 4a. Whereas the upper half of domain D1 and whole of domain D2 show no changes, conformational differences are clearly visible in other portions of flagellin (from F41 (red) to flagellin (blue)): domain D3 moved up slightly; the terminal portions of α -helices in

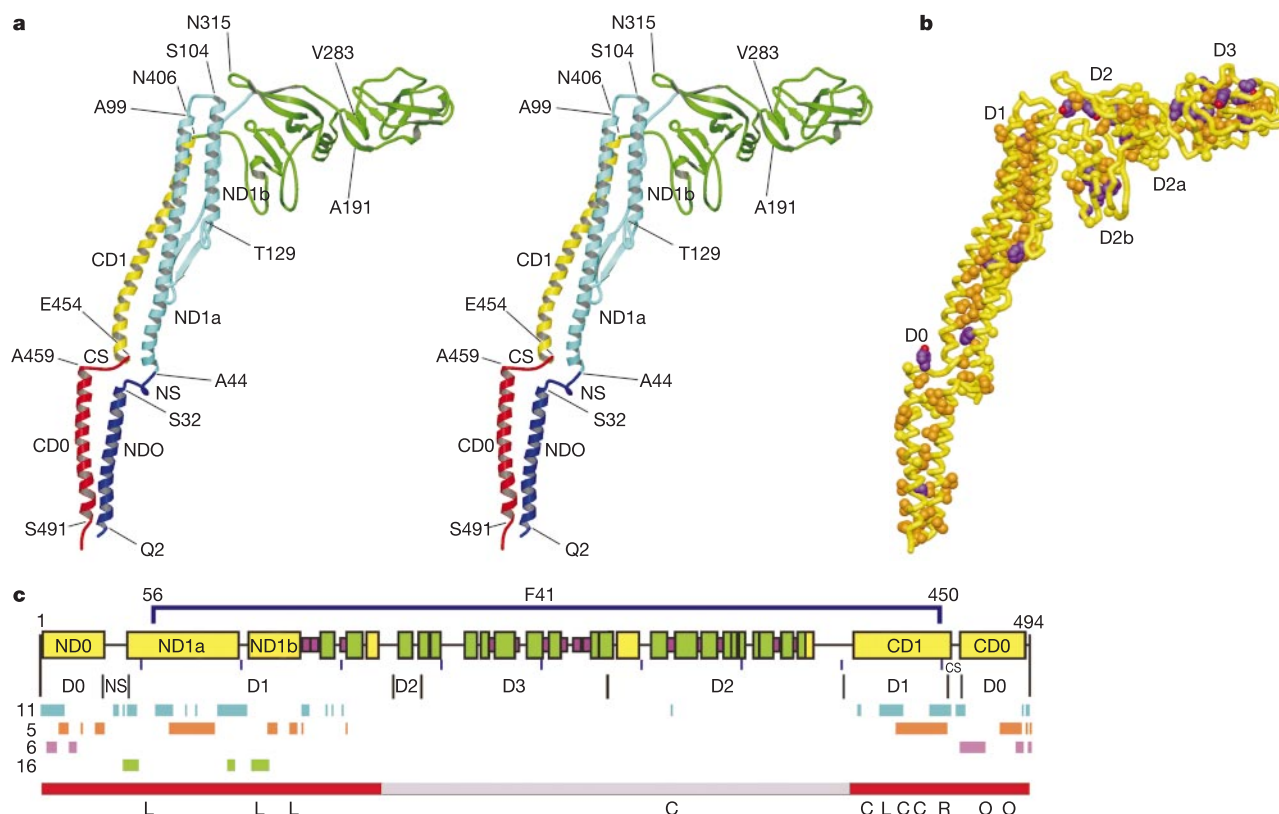


Figure 2 The C α backbone trace, hydrophobic side-chain distribution and structural information of flagellin. **a**, Stereo diagram of the C α backbone. The chain is coloured as follows: residues 1–44, blue; 44–179, cyan; 179–406, green; 406–454, yellow; 454–494, red. **b**, Distribution of hydrophobic side chains, mainly showing hydrophobic cores that define domains D0, D1, D2a, D2b and D3. Side-chain atoms are coloured as follows: Ala and Met, yellow; Leu, Ile and Val, orange; Phe, Tyr and Pro, purple (carbon) and red (oxygen). **c**, Position and region of various structural features in the amino acid sequence of flagellin. Shown are, from top to bottom: the atomic model of F41 in blue; the secondary structure distribution with α -helix in yellow, β -structure in green and β -turn in

purple; tick mark at every 50th residue in blue; domains D0, D1, D2 and D3 and spoke regions NS and CS; the subunit contact regions along the 11-start in cyan, along the 5-start in orange, along the 6-start in pink and along the 16-start in green; the well-conserved amino-acid sequence in red and variable region in violet; point mutations that produce the filament of different supercoils. Letters at the bottom indicate the morphology of mutant filaments: L (F53V, D107E, R124A, R124S, G426A), L-type straight; R (A449V), R-type straight; C (D313Y, A414V, A427V, N433D), curly; O (Q472L, Q481L, Q481S), coiled²¹. Figures of molecular models presented in Figs 2, 3, 4, 5 and 6 are all prepared with MOLSCRIPT⁴⁷ and RASTER3D⁴⁸.

domain D1 (ND1a and CD1) changed the position and orientation of their helix axes and also elongated. The upward displacement of domain D3 has been observed by energy minimization of the F41 protofilament model¹⁹, suggesting that the F41 conformation in the crystal is affected by the crystal-packing interactions.

The most pronounced difference was observed in the orientation of the CD1 helix. The end of this helix becomes more upright in the filament. The end-on view in Fig. 4b shows this difference clearly. As the subunit packing becomes more intimate by this conformational change, it is likely that this change occurs during the assembly process of flagellin to form the filament. This implies that the

conformation of the terminal portions of F41 is closer to that of monomeric flagellin in solution.

Intersubunit interactions in the filament

The long-standing prediction of the subunit packing interactions in the filament from its helical symmetry was that one subunit would interact with all of the six coordinated subunits: two each along the 11-, 5- and 6-start helices. What we see in the filament core structure is slightly different. Only the inner and outer tubes of the filament are displayed in Fig. 5 and viewed from outside and inside. Subunits are numbered along the 1-start helix as described in the legend. The

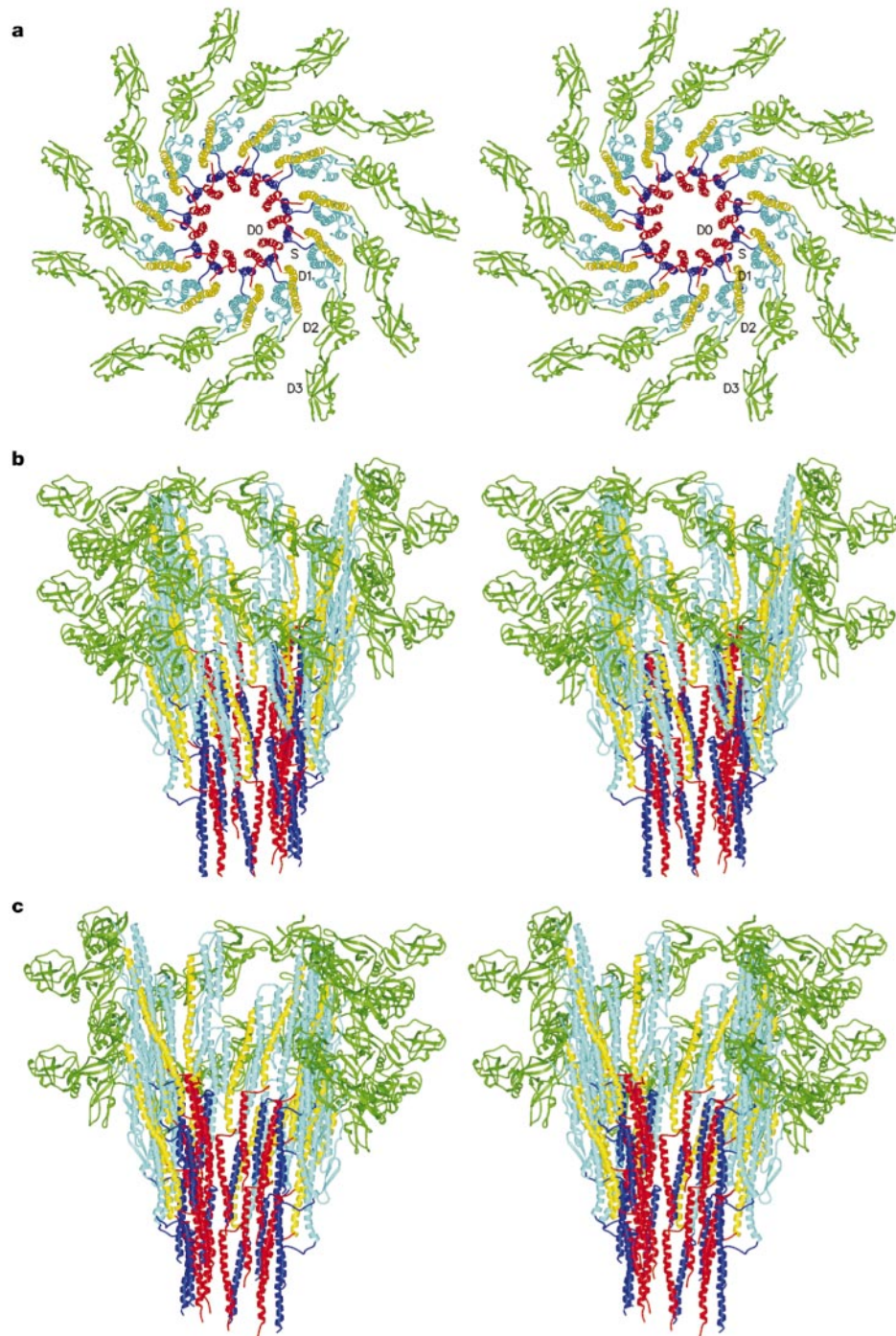


Figure 3 Ribbon diagram of the C α backbone of the filament model in stereo view. **a**, End-on view from the distal end of the filament. Eleven subunits are displayed. **b**, Side view from outside the filament. Three protofilaments on the far side have been removed

for clarity. **c**, Side view from inside the filament. Three protofilaments on the near side have been removed. Top and bottom of the side view images correspond to the distal and proximal ends of the filament, respectively. The chain is colour coded as in Fig. 2a.

protofilaments formed along the 11-start helix are tightly packed laterally with a half-subunit stagger to form the concentric double-tubular structure. In the outer tube, the domains of D1, each of which consists of three α -helices—ND1a (cyan), ND1b (green) and CD1 (yellow)—and a β -hairpin (green), interact with one another along the 11- and 5-start directions only. The interactions along the 11-start helix are similar to those observed in the F41 crystal. Along the 5-start helix, ND1a and ND1b of subunit 0 interact with the β -hairpin of domain D1 and the lower half of CD1 of subunit 5 to form the tight subunit array along the 5-start helix. But, no contact is observed along the 6-start helix; for example, in Fig. 5 subunit 11 intervenes in the space between subunits 0 and 6, or subunit 17 would come into the open space between subunits 6 and 12.

These interactions are more or less the same as have been observed in the F41 filament model (K. Imada *et al.*, manuscript in preparation), which actually simplified the mechanism of twist-curvature coupling for flagellar supercoiling described in the previous study¹³. Now the two switches coupled are a conformational switch of β -hairpin in domain D1 for the curvature¹⁹ and a 2.6 Å displacement of the entire domain D1 along the protofilament relative to its 5-start neighbour for the twist¹³.

We also found the following in our study. There are contacts along the 16-start, for example, between subunit -5 and 11, where direct contacts are present between the top of ND1a and ND1b of

subunit -5 and the end of the NS spoke and the beginning (bottom) of ND1a of subunit 11 (Fig. 5a; near the cluster of hydrophobic residues Phe 53, Phe 131 and Val 449, shown as a space-filling model). A few charge interactions are observed to stabilize the contacts. There are also new contacts along the 11-start, between the middle portion of CD1 of subunit 0 and the end of the

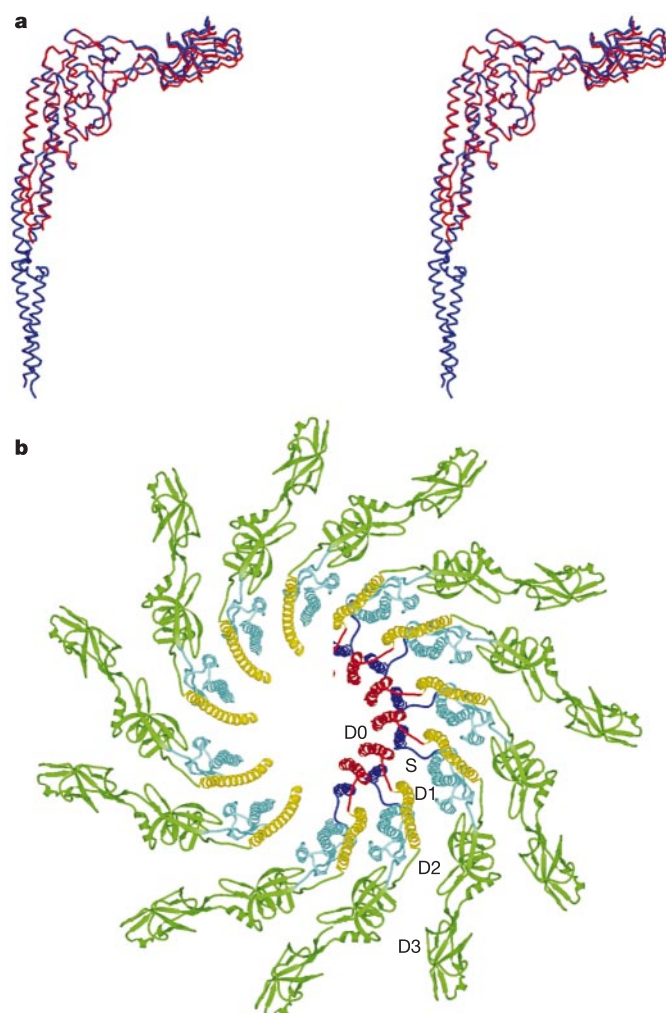


Figure 4 Comparison of the C α backbone of flagellin in the filament with F41 in the crystal. **a**, Stereo diagram of flagellin (blue) and F41 (red). **b**, End-on view of the filament model made of F41 (left half) and flagellin (right half). The chain is colour coded as in Fig. 2a.

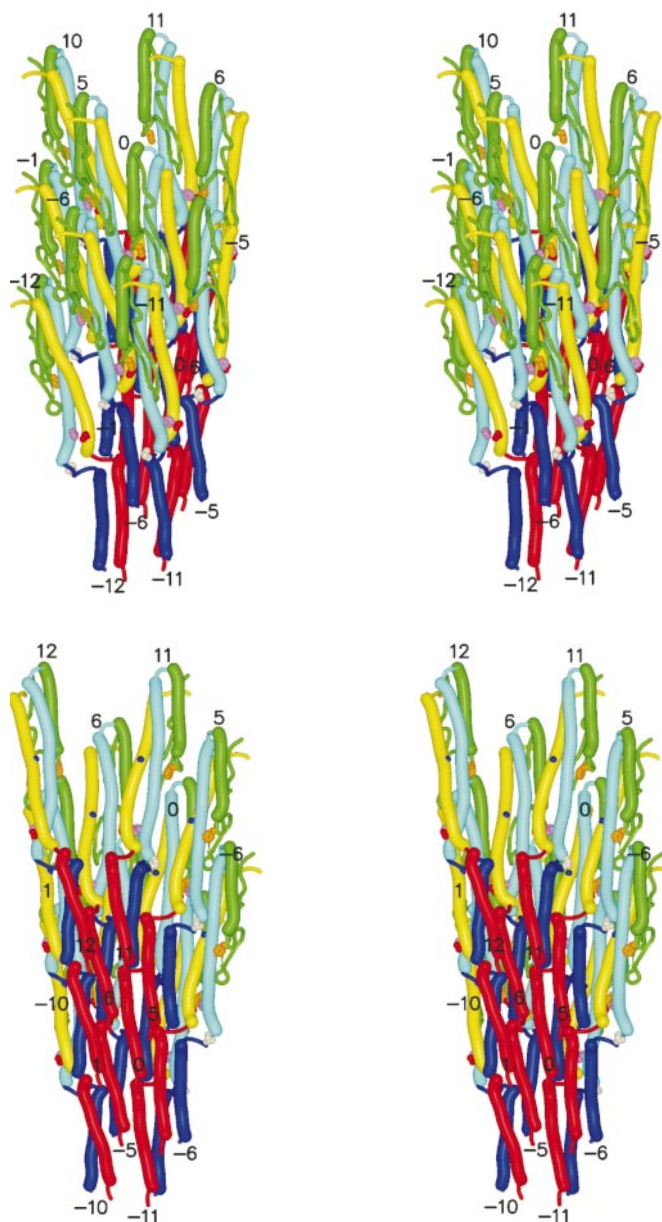


Figure 5 Intersubunit interactions in the inner and outer tubes of the filament in stereo. **a**, Side view from outside the filament. **b**, Side view from inside the filament. Domains D2 and D3 have been removed for clarity. Ten subunits in four protofilaments are displayed, where each protofilament is made of either two or three subunits. Subunits are numbered along the 1-start (right-handed single stranded) helix with a reference subunit numbered 0 and subunits in the distal and proximal directions are given positive and negative numbers, respectively. Various helical lines can be identified easily from these numbers (for example, a helical line connecting -11, 0, 11 is the 11-start). Thick tubes represent α -helices. The chain is coloured as follows: NDO and NS, blue; ND1a, cyan; ND1b and β -hairpin, green; CD1, yellow; CS and CDO, red. See Fig. 2 for labels of α -helices and spokes. Side chains involved in the interactions with amino acid residues that undergo point mutation causing the formation of the R-type straight filament (A449V) and the L-type straight filament (G426A) are presented by space-filling representation. These residues are coloured as follows: Val 449, red; Phe 131, orange; Phe 53, violet; Gly 426, blue; Ile 37, beige.

NS spoke of subunit 11, which obliquely traverses the surface of CD1 (Fig. 5b; marked by a blue dot representing Gly 426). These two interactions are formed by the extension of ND1a towards the proximal direction in the filament compared with F41, and are therefore found neither in the F41 filament model nor in the protofilament structure in the F41 crystal. Although much less extensive compared with the other interactions, these interactions seem to have important roles in the polymorphic supercoiling mechanism as discussed later.

In the inner tube there are extensive and intricate interactions between the domains of D0 in all directions of three major helices, 5-, 6- and 11-start, but not in the direction of 16-start. Domain D0 consists of two α -helices, ND0 (blue) and CD0 (red), as shown in Fig. 5. Along the 11-start, the open end at the bottom of the α -helical coiled coil formed by ND0 and CD0 of the upper subunit (11) caps the end of CS and the top of CD0 of the lower subunit (0), where Tyr 458 (Fig. 1) sticks up from CS into the hydrophobic pocket of the open end (Fig. 5b). In the same pair, the N-terminal portion (bottom) of ND0 runs up along the bottom of CD1 in an antiparallel manner. This is the only major contact between the inner and outer tubes, which are basically well separated from each other by the two spoke chains. ND0 α -helices interact with one another along the 5-start (for example, upper half of subunit 0 and

lower half of subunit 5 (or -11 and -6)) but not along the 6-start (for example, -11 and -5) (Fig. 5a). CD0 α -helices make contacts with one another along the 6-start (upper half of subunit 0 and lower half of subunit 6), but not directly along the 5-start (0 and 5) (Fig. 5b). There are contacts along the 6-start between the upper half of CD0 of subunit 0 and the lower half of ND0 and CD0 in subunit 6. There are also minor contacts along the 5-start between the corner at the end of ND0 turning to the NS spoke of subunit 0 and the middle portion of CD0 of subunit 5 (Fig. 5b). Thus, the N- and C-terminal chains of flagellin form intricate intersubunit interactions in the filament structure.

Most of the intersubunit interactions found within the outer tube are polar-polar or charge-polar, and contributions of hydrophobic interactions are relatively small, whereas those found within the inner tube and between the inner and outer tubes are mostly hydrophobic, contributing to the high stability of the filament structure. These features are presented in Fig. 6a–c. By removing three protofilaments on the front face of the filament model, the side-edge surface of the protofilament involved in the lateral interactions is visualized. The hydrophobic patches running along the central axis of domain D1 show their hydrophobic core exposed on the surface but they are not directly involved in the lateral interactions, whereas those showing the prominent hydrophobic features on the edge surface of domain D0 are extensively involved in the lateral interface as described above. Many charged residues are exposed on the surface of domain D1 and the spoke region, whereas only a few are exposed in domain D0.

It was suggested by heptad repeats of hydrophobic amino acid residues in the terminal regions of flagellin as well as other axial components of bacterial flagella that the terminal chains of neighbouring subunits fold together into α -helical coiled coils²⁷. This interlocking organization has been postulated to be the common motif by which the flagellar axial proteins form a continuous, mechanically stable structure. Also, for formation of various filament structures by self-assembly of biological macromolecules, α -helical coiled coil seems to be used often. The needle structures of the type III protein export system of pathogenic bacteria used to inject virulence proteins into host cells²⁸, for example, are thought to have α -helical coiled coil as a common motif²⁹. At least for flagellin, however, we now see that an α -helical coiled coil is formed within each subunit, and it would probably be so for other flagellar and type III axial proteins. But, the prediction was in part correct in the sense that intersubunit hydrophobic interactions are the major force that stabilizes the flagellar filament structure in aqueous solution (see also Supplementary Figs 2 and 3 for the intersubunit interactions).

Central channel

During the growth of the filament, a large number of flagellin molecules, together with a small number of hook-associated proteins—HAP1, HAP2 and HAP3—are synthesized in the cell and are selectively translocated by the flagellar type III protein export system into the long, narrow central channel of the flagellum^{30,31}, through which they are transported to the distal end. The diameter of the central channel was previously estimated to be about 30 Å from the density maps of the filaments at around 10 Å resolution^{14,15}, but in the atomic model it is only about 20 Å (Fig. 6d). Main-chain and side-chain atoms near the C terminus stick out into the channel space, making its diameter significantly smaller. It was thought that domain rearrangements and partial unfolding may be sufficient to allow flagellin molecules to pass through the channel, but it now seems that more considerable unfolding may be required. At least this narrow channel would prevent unfolded flagellar proteins from undesirable aggregation, and much wider space available at the distal end—which is formed by the outer tube of the filament and the HAP2 pentamer cap complex bound to cover the open end of the filament—would function as a folding chamber or Anfinsen's cage³² (Fig. 6c).

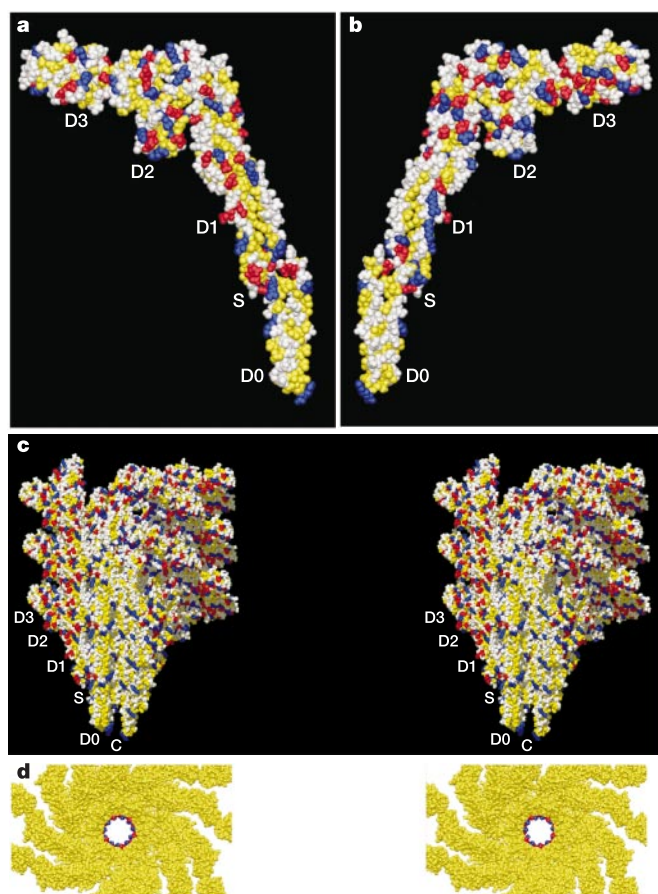


Figure 6 Distribution of charged, polar and hydrophobic residues in space-filling representation. **a, b**, Two side views of a single flagellin molecule. **c**, Side view of the filament in stereo. Three protofilaments on the near side of the filament are removed to show the surface of the protofilament edge and the central channel (labelled C at the bottom). Amino acid residues are colour coded: positively charged, blue; negatively charged, red; polar, white; hydrophobic, yellow. Note that the inner surface of the channel is mainly white (polar), and the edge surface of the outer tube is rich in blue and red colours (charged), but that of the inner tube is mostly yellow (hydrophobic). **d**, End-on view of the filament in stereo from the distal end. The central channel is magnified. Three residues, Gln 484 (red), Asn 488 (red) and Arg 494 (blue), stick out into the channel.

The inner surface of the channel consists of mainly polar amino acids with one positively charged residue, Arg 494 (Fig. 6d). An acidic staining reagent, uranyl acetate, used for electron microscopy cannot stain the central channel, probably because the positive charges on the inner surface repel it. The polar nature of the surface may be advantageous for fast diffusion of unfolded proteins, because unfolded proteins should have many hydrophobic side chains exposed, which would be trapped on the channel surface of hydrophobic nature. We predict that this is also a common feature of the type III protein export system including the needle complex of pathogenic bacteria.

Mutations that affect the supercoiling

About a dozen point mutation sites have been identified for a wild-type strain of *S. typhimurium* SJW1103 in relation to the filament morphology²¹, and all except one are found within highly conserved terminal regions—approximately 170 residues from the N terminus and about 90 residues from the C terminus—which form the core of the filament as shown in Figs 2 and 3. With a complete atomic model of the filament, we can now make more comprehensible interpretations of some of these mutations.

Mutation A449V (Fig. 5b) is responsible for the formation of the R-type straight filament^{20,21}, on which we carried out the structure analysis reported here. Amino acid residue 449 is located at the bottom of CD1, which forms an α -helical coiled coil with ND1a, and it is surrounded by Phe 53 in ND1a of the same subunit and Phe 131 in the first of two consecutive β -turns just after ND1b of the 5-start neighbour on the proximal side (Fig. 5a). The β -hairpin after these β -turns was identified to be responsible for the conformational switching for the curvature formation¹⁹. The axial displacement of domain D1 relative to its 5-start neighbour by 2.6 Å along the protofilament is the essential feature responsible for the change of twist¹³. The mutation from alanine to valine at 449 obviously stabilizes its hydrophobic interactions with two surrounding phenylalanine side chains, and therefore stabilizes the lateral disposition of the protofilaments in the R-type structure. This also explains why flagellin with mutation F53V forms the L-type straight filament. The reduction of the hydrophobic interactions in this region destabilizes the R-type lateral disposition of the protofilaments.

Mutation G426A (Fig. 5b) is responsible for the formation of the L-type straight filament^{20,21}. Residue 426 is located in the middle of CD1 (blue dot in Fig. 5b) and its C α atom directly faces the carbonyl oxygen of Ile 37 in the NS spoke of the upper subunit in the protofilament, where the NS chain obliquely traverses the upper surface of CD1. When mutated from glycine to alanine, the C β atom of alanine would push up NS at the carbonyl oxygen and make the repeat distance along the protofilament slightly longer. As the repeat distance along the L-type protofilament is longer than that of the R-type by 0.8 Å, the structural change by the mutation G426A would stabilize the L-type structure.

Atomic model by image analysis

The atomic model that we present here is the first one that has been built on a three-dimensional density map obtained by electron microscopy and image analysis alone. There have been a few cases of macromolecular structure analyses by electron crystallography at near atomic resolutions where density maps allowed atomic models to be built^{33–36}, but these analyses were all carried out by using electron diffraction as well as image analysis of relatively large two-dimensional protein crystals. Electron diffraction was essential to provide sufficiently high signal to noise ratio in the structure amplitude data, because the amplitude data are rather poorly obtained by image analysis of electron cryomicrographs. Here, we have demonstrated that a three-dimensional density map of macromolecular structures can be obtained at a resolution more or less sufficient to build an atomic model by careful image analysis of a relatively small number of molecular images (about 40,000). The

number of molecular images required to achieve 3 Å resolution has been estimated to be about 12,000, provided that images of ideal quality are used³⁷. Although the number of images we used here is still far larger than this estimation, it is much smaller than those that have actually been used for the structure analysis of two-dimensional protein crystals (a few to several million)^{33–36}.

For single-particle image analysis, it is generally believed that the number of images from a few tens of thousands to hundreds of thousands is required to obtain a map at 10–7 Å resolution. The reason why a relatively small number of images allowed us to obtain a high-resolution map can be considered as follows. First, a liquid-helium-cooled and highly stable specimen stage of the electron microscope³⁸ produced high-quality images with sufficient reduction of radiation damage in high-resolution structural information. Second, the helical image reconstruction method with many newly developed programs and algorithms allowed accurate alignment of each image (see Methods). Especially, solvent flattening of individual images seems to be very powerful in removing the noise from solvent regions, making the image alignment more accurate, and increasing the resolution that can be obtained by a given number of images, as discussed previously²³. Third, the structural order and helical symmetry of the flagellar filament is high enough to give rise to high-resolution structural data, which is demonstrated by X-ray fibre diffraction patterns showing sharp layer-line reflections beyond 3 Å resolution¹³. The atomic model of F41 (ref. 19) was, of course, useful not only to check the reliability of the present density map but also to build the complete atomic model of the filament.

It is useful for those working in the field of biological and medical sciences to be able to look at the three-dimensional atomic arrangements of macromolecules and molecular assemblies without making crystals, not only because many of them are hard to crystallize but also because crystal packing would affect the structures, as seen in this work, and therefore limit the functional states that we can visualize. NMR spectroscopy, which permits structure analysis in solution, still has a limitation in the molecular mass of macromolecules to be analysed: the upper limit being a few tens of kilodaltons. Electron cryomicroscopy is a potentially powerful method in this regard because it can be applied to various forms of samples; that is, two-dimensional crystal, helical assembly, spherical assembly and even single molecule, although attainable resolution depends on the system. The present work suggests that even single-particle image analysis at near atomic resolution may not be unrealistic any more, provided that highly accurate alignment becomes possible.

Towards the mechanism of supercoiling

The structure of the R-type straight filament has provided many insights into the molecular interactions within the filament, flagellin transport and mutations responsible for polymorphic supercoiling. However, the true understanding of the mechanism of polymorphic supercoiling and its dynamic transition has to wait until the atomic model of the L-type straight filament becomes available. The structure analysis is now underway (S.M.-Y., K.Y. and K.N., manuscript in preparation). Also, to visualize the conformational change for the switch mechanism, molecular dynamics simulations using a short filament model with 2.4 million atoms including solvent molecules are now under way (A. Kitao *et al.*, manuscript in preparation). □

Methods

Sample preparation

Flagellin was isolated from *S. typhimurium* strain SJW1655 and the R-type straight flagellar filament was reconstituted as previously described¹⁴.

Electron cryomicroscopy

Images of frozen hydrated filaments were taken with a JEOL JEM-3000SFF electron

microscope with its field emission gun operated at 300 kV and the specimen temperature of 4 K. Images were recorded on SO-163 film (Eastman Kodak Co.) at a magnification of $\times 50,000$. The electron dose was about $20 \text{ e}^- \text{Å}^{-2}$.

Image analysis

Images were examined first by optical diffraction and digitized with LeafScan 45 (Scitex) at a step size of 5 μm , which corresponds to 1 Å on the sample plane. Helical image reconstruction was carried out as previously described^{23,39} with some modifications. Briefly, newly developed GUI programs (K.Y., C. Toyoshima, S.M.-Y. and K.N., manuscript in preparation) were introduced to speed up the image processing of the individual filaments, such as determining the box parameters including the repeat distance and out-of-plane tilt³⁹. Then, the box parameters were further refined to get sharper layer-lines and amplitudes of higher quality in the Fourier transform (K.Y. and C. Toyoshima, unpublished data). Three-dimensional distortions of the filaments were partially corrected by dividing a filament image into several segments and fitting them in the reciprocal space to the reference image as described previously²² with slight modification (K.Y. & C. Toyoshima, unpublished data). The images were corrected for the CTF by applying the curvature of the Ewald sphere⁴⁰. Then, solvent flattening was applied to individual images to remove the noise in the solvent region and that convoluted from the solvent region into the filament image owing to the CTF²³. The number of filament images used for the final reconstruction was 102, and the total number of molecular images was 41,469. A more detailed description of the image analysis will be given elsewhere.

Model building

The atomic model of flagellin fragment F41 (ref. 19) was fitted to the density map using O⁴¹. Then, an initial model of full-length flagellin was built by tracing missing terminal chains. The model was refined using both positional and simulated annealing refinements⁴² by a molecular dynamics refinement program, FEX-PLOR, which we developed based on FX-PLOR²⁴ for electron microscope image analysis of the helical assembly. The amplitude-weighted phase-residual was implemented in FEX-PLOR as an effective potential energy. The layer-line amplitude distributions of the electron microscope data were then scaled to the structure factors calculated from the model based on their radial amplitude profiles obtained by averaging the amplitudes within each resolution shell. The density map was calculated again, and model building and refinement were iterated. We used PROCHECK⁴³ to obtain a Ramachandran plot for the final model including all 494 residues of flagellin. A total of 89% of the residues fell into the most favoured regions, whereas no residues in the inner domains (domain D0, D1 and the spoke region) fell into generously allowed or disallowed regions. The quality of the map as well as the fitting of the atomic model was checked by the FOM obtained by combining the phases (equation (1) in the legend to Supplementary Table 2). Combination of phases from the model and observed images was carried out by equation (2) in the legend to Supplementary Table 2. The definition is similar to that used in solvent flattening or molecular averaging in X-ray crystallography⁴⁴. A more detailed description of model building and refinement will be given elsewhere.

We divided the Fourier space into three fan-shaped sectors, I, II and III. Sector I is close to the equator, II in the middle, and III close to the meridian. An FOM of 0.45 is equivalent to a phase error of approximately 63° and this value was used as the criterion to determine the resolution limit. The FOM decreased as the resolution became higher, but it was higher than 0.45 out to 5 Å resolution in all directions. In sector I the FOM decreased to 0.45 and became flat beyond 5 Å resolution, but in sector III it went on with a value above 0.45 out to about 4 Å. This is how we determined the resolution limit.

Received 7 March; accepted 21 May 2003; doi:10.1038/nature01830.

- Berg, H. C. & Anderson, R. A. Bacteria swim by rotating their flagellar filaments. *Nature* **245**, 380–382 (1973).
- Silverman, M. & Simon, M. Flagellar rotation and the mechanism of bacterial motility. *Nature* **249**, 73–74 (1974).
- Kudo, S., Magariyama, Y. & Aizawa, S.-I. Abrupt changes in flagellar rotation observed by laser darkfield microscopy. *Nature* **346**, 677–680 (1990).
- Ryu, W. S., Berry, R. M. & Berg, H. C. Torque-generating units of the flagellar motor of *Escherichia coli* have a high duty ratio. *Nature* **403**, 444–447 (2000).
- Larsen, S. H., Reader, R. W., Kort, E. N., Tso, W. W. & Adler, J. Change in direction of flagellar rotation is the basis of the chemotactic response in *Escherichia coli*. *Nature* **249**, 74–77 (1974).
- Macnab, R. M. & Ornston, M. K. Normal-to-curly flagellar transitions and their role in bacterial tumbling. Stabilization of an alternative quaternary structure by mechanical force. *J. Mol. Biol.* **112**, 1–30 (1977).
- Turner, L., Ryu, W. S. & Berg, H. C. Real-time imaging of fluorescent flagellar filaments. *J. Bacteriol.* **182**, 2793–2801 (2000).
- O'Brien, E. J. & Bennett, P. M. Structure of straight flagella from a mutant *Salmonella*. *J. Mol. Biol.* **70**, 133–152 (1972).
- Asakura, S. Polymerization of flagellin and polymorphism of flagella. *Adv. Biophys.* **1**, 99–155 (1970).
- Calladine, C. R. Construction of bacterial flagella. *Nature* **225**, 121–124 (1975).
- Calladine, C. R. Design requirements for the construction of bacterial flagella. *J. Theor. Biol.* **57**, 469–489 (1976).
- Calladine, C. R. Change of waveform in bacterial flagella: The role of mechanics at the molecular level. *J. Mol. Biol.* **118**, 457–479 (1978).
- Yamashita, I. *et al.* Structure and switching of bacterial flagellar filament studied by X-ray fiber diffraction. *Nature Struct. Biol.* **5**, 125–132 (1998).
- Mimori, Y. *et al.* The structure of the R-type straight flagellar filament of *Salmonella* at 9 Å resolution by electron cryomicroscopy. *J. Mol. Biol.* **249**, 69–87 (1995).
- Morgan, D. G., Owen, C., Melanson, L. A. & DeRosier, D. J. Structure of bacterial flagellar filaments at 11 Å resolution: Packing of the α -helices. *J. Mol. Biol.* **249**, 88–110 (1995).
- Mimori-Kiyosue, Y., Vonderviszt, F., Yamashita, I., Fujiyoshi, Y. & Namba, K. Direct interaction of

- flagellin termini essential for polymorphic ability of flagellar filament. *Proc. Natl Acad. Sci. USA* **93**, 15108–15113 (1996).
- Mimori-Kiyosue, Y., Vonderviszt, F. & Namba, K. Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism. *J. Mol. Biol.* **270**, 222–237 (1997).
- Mimori-Kiyosue, Y., Yamashita, I., Fujiyoshi, Y., Yamaguchi, S. & Namba, K. Role of the outermost subdomain of *Salmonella* flagellin in the filament structure revealed by electron cryomicroscopy. *J. Mol. Biol.* **284**, 521–530 (1998).
- Samatey, F. A. *et al.* Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling. *Nature* **410**, 331–337 (2001).
- Hyman, H. C. & Trachtenberg, S. Point mutations that lock *Salmonella typhimurium* flagellar filaments in the straight right-handed and left-handed forms and their relation to filament superhelicity. *J. Mol. Biol.* **220**, 79–88 (1991).
- Kanto, S., Okino, H., Aizawa, S.-I. & Yamaguchi, S. Amino acids responsible for flagellar shape are distributed in terminal regions of flagellin. *J. Mol. Biol.* **219**, 471–480 (1991).
- Beroukhim, R. & Unwin, N. Distortion correction of tubular crystals: improvements in the acetylcholine receptor structure. *Ultramicroscopy* **70**, 57–81 (1997).
- Yonekura, K. & Toyoshima, C. Structure determination of tubular crystals of membrane proteins. III. Solvent flattening. *Ultramicroscopy* **84**, 29–45 (2000).
- Wang, H. & Stubbs, G. Molecular dynamics in refinement against fiber diffraction data. *Acta Crystallogr. A* **49**, 504–513 (1993).
- Maki, S., Imada, K., Furukawa, Y., Vonderviszt, F. & Namba, K. Plugging interactions of HAP2 pentamer into the distal end of flagellar filament revealed by electron microscopy. *J. Mol. Biol.* **277**, 771–777 (1998).
- Vonderviszt, F., Kanto, S., Aizawa, S.-I. & Namba, K. Terminal region of flagellin are disordered in solution. *J. Mol. Biol.* **209**, 127–133 (1989).
- Homma, M., DeRosier, D. J. & Macnab, R. M. Flagellar hook and hook-associated proteins of *Salmonella typhimurium* and their relationship to other axial components of the flagellum. *J. Mol. Biol.* **213**, 819–832 (1990).
- Kubori, T. *et al.* Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system. *Science* **280**, 602–605 (1998).
- Delahay, R. M. & Frankel, G. Coiled-coil proteins associated with type III secretion systems: a versatile domain revisited. *Mol. Microbiol.* **45**, 905–916 (2002).
- Fan, F., Ohnishi, K., Francis, N. R. & Macnab, R. M. The FliP and FliR proteins of *Salmonella typhimurium*, putative components of the type III flagellar export apparatus, are located in the flagellar basal body. *Mol. Microbiol.* **26**, 1035–1046 (1997).
- Minamino, T. & Macnab, R. M. Components of the *Salmonella* flagellar export apparatus and classification of export substrates. *J. Bacteriol.* **181**, 1388–1394 (1999).
- Yonekura, K. *et al.* The bacterial flagellar cap as the rotary promoter of flagellin self-assembly. *Science* **290**, 2148–2152 (2000).
- Henderson, R. *et al.* An atomic model for the structure of bacteriorhodopsin. *J. Mol. Biol.* **213**, 899–929 (1990).
- Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367**, 614–621 (1994).
- Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
- Nogales, E., Sharon, G. W. & Downing, K. H. Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature* **391**, 199–203 (1998).
- Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).
- Fujiyoshi, Y. The structural study of membrane proteins by electron crystallography. *Adv. Biophys.* **35**, 25–80 (1998).
- Toyoshima, C. Structure determination of tubular crystals of membrane proteins. I. Indexing of diffraction patterns. *Ultramicroscopy* **84**, 1–14 (2000).
- DeRosier, D. J. Correction of high-resolution data for curvature of the Ewald sphere. *Ultramicroscopy* **81**, 83–98 (2000).
- Jones, T. A., Zhou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallography R factor refinement by molecular dynamics. *Science* **235**, 458–460 (1987).
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemistry of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
- Drenth, J. *Principles of Protein X-ray Crystallography* (Springer, New York, 1994).
- McRee, D. E. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
- Merritt, E. A. & Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank Y. Fujiyoshi for technical advice on the use of the electron cryomicroscope. We also thank F. Oosawa, S. Asakura and D. L. D. Caspar for continuous support and encouragement.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.N. (keiichi@fbs.osaka-u.ac.jp). Atomic coordinates have been deposited in the Protein Data Bank under accession code 1UCU.

An asymptotic-giant-branch star in the progenitor system of a type Ia supernova

Mario Hamuy¹, M. M. Phillips², Nicholas B. Suntzeff³, José Maza⁴, L. E. González⁴, Miguel Roth², Kevin Krisciunas², Nidia Morrell², E. M. Green⁵, S. E. Persson¹ & P. J. McCarthy¹

¹Carnegie Observatories, 813 Santa Barbara Street, Pasadena, California 91101, USA

²Las Campanas Observatory, Carnegie Observatories, Casilla 601, and

³Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatories, Casilla 603, La Serena, Chile

⁴Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile

⁵University of Arizona, Steward Observatory, Tucson, Arizona 85721, USA

Stars that explode as supernovae come in two main classes. A type Ia supernova is recognized by the absence of hydrogen and the presence of elements such as silicon and sulphur in its spectrum; this class of supernova is thought to produce the majority of iron-peak elements in the Universe. They are also used as precise 'standard candles' to measure the distances to galaxies. While there is general agreement that a type Ia supernova is produced by an exploding white dwarf star¹, no progenitor system has ever been directly observed. Significant effort has gone into searching for circumstellar material to help discriminate between the possible kinds of progenitor systems², but no such material has hitherto been found associated with a type Ia supernova³. Here we report the presence of strong hydrogen emission associated with the type Ia supernova SN2002ic, indicating the presence of large amounts of circumstellar material. We infer from this that the progenitor system contained a massive asymptotic-giant-branch star that lost several solar masses of hydrogen-rich gas before the supernova explosion.

SN2002ic was discovered before or near maximum light by the Nearby Supernova Factory search⁴. It appeared ~4 arcsec west of a faint elongated galaxy and north of a fainter galaxy (see Supplementary Fig. 1). We have measured a redshift of $z = 0.22$ for the galaxy to the east and $z = 0.078$ for the galaxy to the south, which rules out any association of these galaxies with the supernova ($z = 0.0666$; see below).

Our spectroscopic coverage of SN2002ic encompasses a period of ~60 days, and clearly demonstrates that this object belongs to the Ia class (see Fig. 1). The features in SN2002ic are very much like those of the type Ia SN1991T⁵, but diluted in strength. Although SN1991T/1999aa-like events are sometimes referred to as 'peculiar', they make up 20% of the local population of type Ia supernovae (SNe Ia)⁶. The most remarkable spectroscopic feature in SN2002ic is the H α emission at $z = 0.0666$. This emission is spatially unresolved with a full-width at half-maximum, FWHM, of ≤ 1.2 arcsec (1.7 kpc for $H_0 = 65 \text{ km s}^{-1} \text{ Mpc}^{-1}$). The H α profiles (Fig. 2) have an unresolved (FWHM $< 300 \text{ km s}^{-1}$) component on top of a broad (FWHM $\approx 1,800 \text{ km s}^{-1}$) base. While the narrow component could be an unrelated H II region, the broader component cannot be explained in this way. Such narrow + broad H α profiles have no precedent among SNe Ia³, but are typical of SNe IIn⁷ where it is thought that the narrow lines arise from the un-shocked circumstellar material (CSM) photoionized by radiation from the SN, while the intermediate width lines are formed in the shock-heated CSM^{8,9}. The origin of the CSM in SNe IIn is attributed to the SN progenitor, a presumably massive star that undergoes mass loss before explosion. By analogy, the double component H α profile in SN2002ic is clear evidence for a dense CSM. The measured H α

fluxes in SN2002ic correspond to energies of $\sim 2 \times 10^{40} \text{ erg s}^{-1}$ in each of the broad and narrow line components. Photoionization models for the narrow component³ imply an unexpectedly high mass-loss rate of $\sim 10^{-2.4} M_{\odot} \text{ yr}^{-1}$ (where $M_{\odot}$ is the solar mass). The absence of a rapid dimming in H α or a change in linewidth (Fig. 2) suggests that the SN/CSM interaction remains strong at least 2 months after explosion, which is also consistent with large mass-loss rates³.

The SN/CSM interaction in SNe IIn produces a luminosity enhancement compared to classical SNe II¹⁰ owing to the conversion of kinetic energy into continuum radiation. This suggests that the absorption features and light curves of SN2002ic may be similarly 'veiled' relative to 'normal' SNe Ia. This hypothesis is confirmed from the spectroscopic comparison between SN2002ic and the type Ia SN1999ee¹¹ (Fig. 3), and the photometric analysis shown in Fig. 4.

The strength of the SN/CSM interaction in SN2002ic is totally unexpected for a SN Ia, but is typical in SNe IIn. It is interesting to hypothesize that some SNe IIn may actually be SNe Ia like SN2002ic,

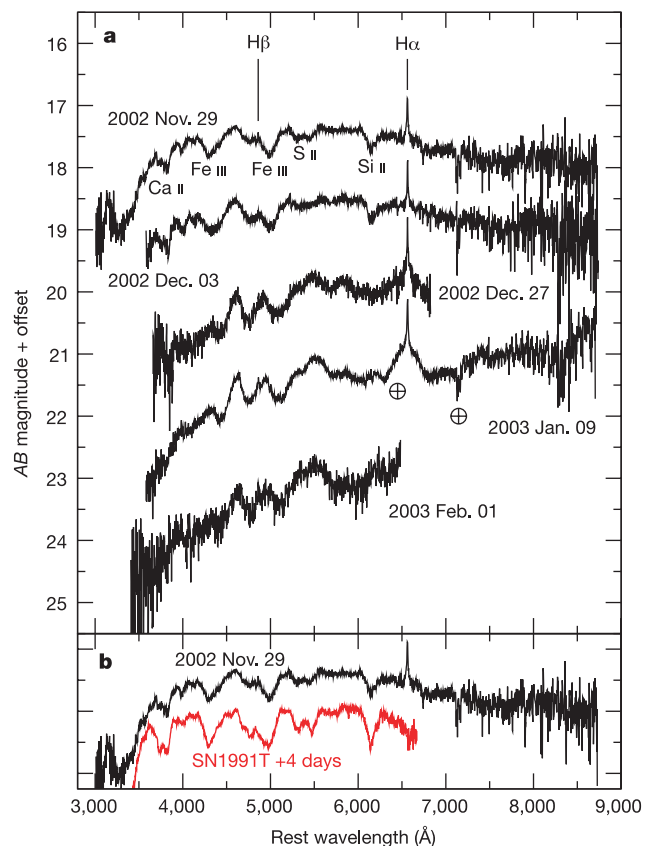


Figure 1 Spectroscopic evolution of SN2002ic. **a**, This sequence shows five spectra of SN2002ic (in AB magnitudes) obtained between 2002 November 29 and 2003 February 1 UT with the Las Campanas Observatory Baade 6.5-m and du Pont 2.5-m telescopes, and the Steward Observatory Bok 2.3-m telescope. Arbitrary offsets have been added to the spectra for clarity. The spectra are +6, +10, +34, +47 and +70 days from estimated maximum light. We attempted to remove the two most prominent telluric lines (indicated with the crossed circles), but some residuals are evident. The top spectrum shows the Si II $\lambda 6,355$ feature that defines the Ia class, as well as prominent Fe III absorption features at 4,200 and 4,900 Å. The absence of the He/Na feature at 5,900 Å in the spectral evolution rules out a type Ib/c classification. **b**, A comparison between the 2002 November 29 (+6 days) observation of SN2002ic and the spectrum of the type Ia SN1991T⁵ obtained at an epoch of +4 days, shows that both spectra are quite similar, except that the features in SN2002ic are all diluted in strength.

but with an even stronger SN/CSM interaction. In this context, we draw attention to the type IIn SN1997cy that may have been associated with the γ -ray burst source GRB970514^{12,13}. In Fig. 5 we show that the late-time (~ 70 day) spectra of SN2002ic and SN1997cy are strikingly similar. Both the *BV* light curve decline rates (~ 0.7 mag per 100 days) and the peak absolute magnitude of $M(V) \approx -20.1$ of SN1997cy¹³ are very similar to the decline rates inferred for the SN/CSM interaction and the peak brightness for SN2002ic (see Fig. 4). Note that similar high luminosities and flat, constant-colour light curves were also seen in the SN IIn 1988Z¹⁴.

The photometric properties of SN1997cy are well reproduced by a model of the explosion of a $25 M_{\odot}$ star with high explosion energy (3×10^{52} erg) which interacts with a dense CSM of $\sim 5 M_{\odot}$ (ref. 13). In this model, the light curve is powered by the SN/CSM interaction with a contribution of radioactive heating of up to $0.7 M_{\odot}$ of ^{56}Ni . In view of the close similarity between SN1997cy and SN2002ic, and the clear evidence that SN2002ic was a bona fide SN Ia, a re-examination of possible models for SN1997cy seems worthwhile, including both its association to GRB970514 and the overall energetics. We note that C/O Chandrasekhar white dwarfs are

very unlikely to produce more than 2×10^{51} erg (ref. 15). This upper limit appears to be in conflict with the total radiated energy of 10^{52} erg estimated for the SN IIn 1988Z¹⁶, but this estimate is highly uncertain.

Is there evidence for a SN/CSM interaction in other SNe Ia? It has been observed that the *BVRI* light curves of the 1991T/1999a type⁶ show an 'over-brightness' ~ 1 month after maximum compared to spectroscopically 'normal' SNe Ia^{17,18}. Perhaps all 1991T/1999aa events occur in progenitors with a significant CSM, and the SN/CSM interaction accounts for the additional luminosity at later epochs. A test of this hypothesis would be to search for radio or X-ray emission from these SNe Ia or extremely faint hydrogen emission.

The modelling of SN1997cy shows that a few solar masses of CSM material are involved in the emission processes. By analogy, we would expect a star in the progenitor system of SN2002ic to lose a similar amount of mass, ruling out a double-degenerate model for the progenitor of this type Ia explosion. A progenitor consistent with this amount of mass loss is a binary system containing a C/O white dwarf and a massive ($3\text{--}7 M_{\odot}$) asymptotic-giant-branch star

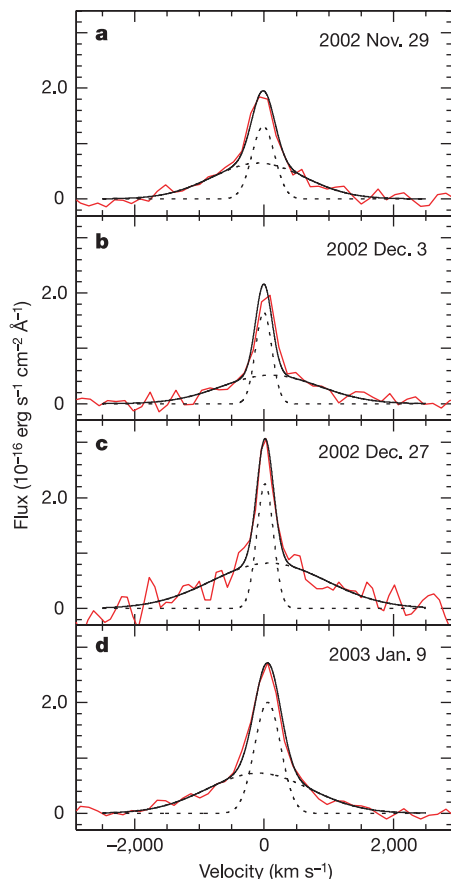


Figure 2 Decomposition and evolution of the $H\alpha$ profiles. The observed $H\alpha$ profiles are shown in red for four different epochs. These profiles cannot be fitted with single gaussian models, but two gaussians of different widths do provide acceptable fits. The black dotted lines show the individual components (with FWHM ~ 300 and $\sim 1,800 \text{ km s}^{-1}$) and the black solid lines the sum of both gaussians. This figure suggests that the $H\alpha$ profile does not evolve rapidly with time. It is difficult to make more precise statements because, while the subtraction of the continuum at early epochs is straightforward, at later times the SN has a broad emission feature (see Fig. 1a) which makes the subtraction much more uncertain.

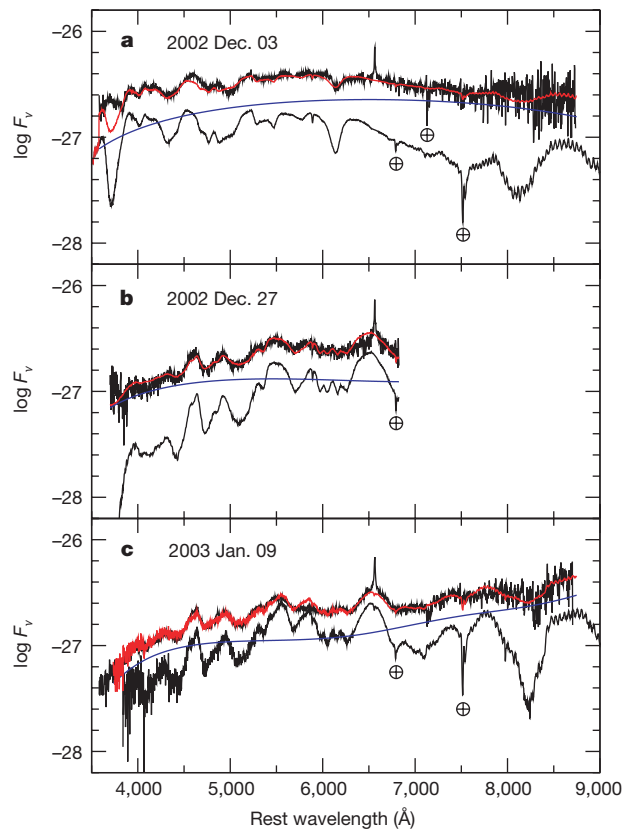


Figure 3 Spectral analysis of SN2002ic. **a**, Spectrum of SN2002ic taken on 2002 December 3 UT (heavy black line) compared with that of the normal type Ia SN 1999ee¹¹ taken 6 days past maximum (thin black line). In red is shown the sum (by eye) of a low-order continuum (blue line) and the spectrum of SN1999ee. The most prominent telluric lines are indicated with crossed circles. **b**, Same as above, except comparing the spectrum of SN2002ic obtained on 2002 December 27 UT with that of SN1999ee taken 34 days past maximum. **c**, Same as above, except comparing the spectrum of SN2002ic taken on 2003 January 9 UT with that of SN1990N obtained 47 days past maximum. Except for the features of Ca II at $3,950 \text{ \AA}$ and $8,600 \text{ \AA}$, which appear weaker in SN2002ic, in all three cases the continuum + SN spectra and the SN2002ic spectra match extremely well. This simple model explains why the absorption features in SN2002ic appear 'veiled' relative to normal SNe Ia.

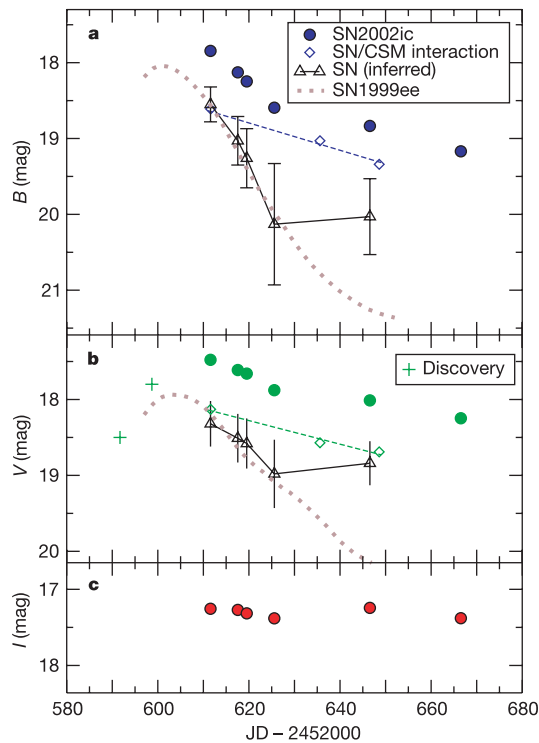


Figure 4 Photometric analysis of SN2002ic. Filled circles, light curves of SN2002ic obtained with the Las Campanas 2.5-m and 1-m telescopes (**a**, B band; **b**, V band; **c**, I band). These light curves were derived from differential measurements relative to several local standards calibrated on two photometric nights. Also shown with the 'plus' symbols in **b** are the unfiltered discovery magnitudes⁴. Maximum light occurred around November 23 (JD 2452602) and the peak magnitudes were $B \approx 17.7$, $V \approx 17.4$ and $I \approx 17.2$. Correcting for a Galactic reddening of $E(B - V) = 0.073$ (ref. 22) and K corrections²³, we derive absolute magnitudes of $M(B) \approx -20.1$, $M(V) \approx -20.3$ and $M(I) \approx -20.4$ ($H_0 = 65$), which are ~ 1 mag brighter than the most luminous SNe Ia²⁴. For comparison are plotted the B and V light curves of the 'normal' type Ia SN1999ee²⁵ (characterized by a post-maximum decline rate, $\Delta m_{15}(B)$, of 0.94) expected for the redshift and reddening of SN2002ic. While the initial decline rate of SN2002ic was very slow, the B and V light curves evolved into the final linear decline rate within ~ 25 days of maximum, which is only observed for the fastest declining SNe Ia²⁶. Also shown (diamonds) in B and V is the luminosity evolution of the SN/CSM interaction as inferred from the dilution of the spectral features (Fig. 3) to which we have fitted straight lines (shown as dashes). The triangles show the result of subtracting this light curve from the observed photometry of SN2002ic. For the first ~ 25 days, the 'unveiled' peak magnitudes of $B \approx V \approx 18.0$ are consistent with a 'normal' SN Ia. However, beyond JD 2452630 the residual flux remains high compared to SN1999ee. This suggests that a model of the light curves of SN2002ic as the sum of a 'normal' SN Ia and a slow-declining SN/CSM interaction may be overly simple.

where the integrated mass loss can reach a few solar masses¹⁹. The accretion of part of this wind onto the white dwarf could bring it to the Chandrasekhar mass. An alternative explanation is the thermonuclear explosion of the degenerate core of a 'single' asymptotic-giant-branch star in a 'type 1.5' event²⁰, although the spectrum of this event has not been calculated in detail. We also note that SN2002ic was much more luminous than its (yet-to-be-identified) host galaxy. Curiously, the host galaxy of SN1997cy was also of very low luminosity¹², as was the host of the 1999aa-like SN1999aw¹⁸. Such low-luminosity galaxies are likely to be characterized by low metallicities²¹. Hence, the presence of SN/CSM interaction in 1991T/1999aa-like events such as SN2002ic may be related to mass loss in a metal-poor environment. □

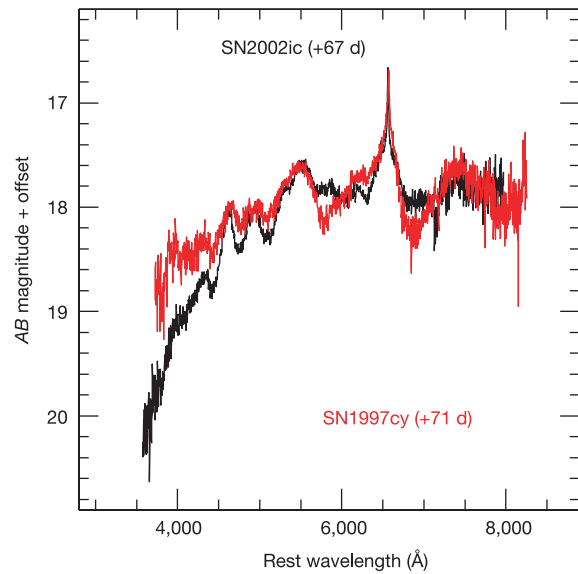


Figure 5 Spectroscopic comparison between SN2002ic and SN1997cy. Spectrum of SN2002ic taken on January 9 (~ 47 days after maximum light, which corresponds to ~ 67 days after explosion for an assumed time of 20 days between explosion and peak brightness) compared to that of the type IIa SN 1997cy taken 71 days after explosion¹³, which is assumed to coincide with the detection of GRB970514¹². The striking similarity between these two objects suggests that some SNe IIa are the result of thermonuclear explosions of white dwarfs surrounded by a dense CSM instead of core collapse in massive stars.

Received 2 April; accepted 16 June 2003; doi:10.1038/nature01854.

1. Livio, M. in *Supernovae and Gamma-Ray Bursts* (eds Livio, M., Panagia, N. & Sahu, K.) 334–335 (STScI Symp. Vol. 13, Cambridge Univ. Press, Cambridge, 2001).
2. Lentz, E. J., Baron, E., Hauschildt, P. H. & Branch, D. Detectability of hydrogen mixing in type Ia supernova premaximum spectra. *Astrophys. J.* **580**, 374–379 (2002).
3. Cumming, R. J., Lundqvist, P., Smith, L. J., Pettini, M. & King, D. Circumstellar H α from SN 1994D and future type Ia supernovae: an observational test of progenitor models. *Mon. Not. R. Astron. Soc.* **283**, 1355–1360 (1996).
4. Wood-Vasey, W. M. *et al.* Supernova 2002ic. *IAU Circ. No.* 8019 (2002).
5. Phillips, M. M. *et al.* SN 1991T: Further evidence of the heterogeneous nature of type Ia supernovae. *Astron. J.* **103**, 1632–1637 (1992).
6. Li, W. *et al.* A high intrinsic peculiarity rate among type Ia supernovae. *Astrophys. J.* **546**, 734–743 (2001).
7. Schlegel, E. M. A new subclass of type II supernovae? *Mon. Not. R. Astron. Soc.* **244**, 269–271 (1990).
8. Chugai, N. N. The origin of supernovae with dense winds. *Astron. Rep.* **41**, 672–681 (1997a).
9. Chugai, N. N. Supernovae in dense winds. *Astrophys. Space Sci.* **252**, 225–236 (1997b).
10. Hamuy, M. in *Core Collapse of Massive Stars* (ed. Fryer, C. L.) (Kluwer, Dordrecht, in the press); preprint at (<http://xxx.lanl.gov/astro-ph/0301006>) (2003).
11. Hamuy, M. *et al.* Optical and infrared spectroscopy of SN 1999ee and SN 1999ex. *Astron. J.* **124**, 417–429 (2002).
12. Germany, L. M., Reiss, D. J., Sadler, E. M., Schmidt, B. P. & Stubbs, C. W. SN 1997cy/GRB 970514: A new piece in the gamma-ray burst puzzle? *Astrophys. J.* **533**, 320–328 (2000).
13. Turatto, M. *et al.* The properties of supernova 1997cy associated with GRB 970514. *Astrophys. J.* **534**, L57–L61 (2000).
14. Turatto, M. *et al.* The type II supernova 1988Z in MCG + 03-28-022: Increasing evidence of interaction of supernova ejecta with a circumstellar ring. *Mon. Not. R. Astron. Soc.* **262**, 128–140 (1993).
15. Arnett, D. *Supernovae and Nucleosynthesis* (Princeton Univ. Press, Princeton, 1996).
16. Aretxaga, I. *et al.* SN 1988Z: Spectro-photometric catalogue and energy estimates. *Mon. Not. R. Astron. Soc.* **309**, 343–354 (1999).
17. Suntzeff, N. B. in *Supernovae and Supernova Remnants* (eds McCray, R. & Wang, Z.) 41 (IAU Colloq. 145, Cambridge Univ. Press, Cambridge, 1996).
18. Strolger, L.-G. *et al.* The type Ia supernova 1999aw: A probable SN 1999aa-like event in a low luminosity host galaxy. *Astron. J.* **124**, 2905–2919 (2002).
19. Weidemann, V. Revision of the initial-to-final mass relation. *Astron. Astrophys.* **363**, 647–656 (2000).
20. Iben, I. & Renzini, A. Asymptotic giant branch evolution and beyond. *Annu. Rev. Astron. Astrophys.* **21**, 271–342 (1983).
21. Henry, R. B. C. & Worthey, G. The distribution of heavy elements in spiral and elliptical galaxies. *Publ. Astron. Soc. Pacif.* **111**, 919–945 (1999).
22. Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
23. Hamuy, M., Phillips, M. M., Wells, L. A. & Maza, J. K. corrections for type Ia supernovae. *Publ. Astron. Soc. Pacif.* **105**, 787–793 (1993).

24. Phillips, M. M. *et al.* The reddening-free decline rate versus luminosity relationship for type Ia supernovae. *Astron. J.* **118**, 1766–1776 (1999).
25. Stritzinger, M. *et al.* Optical photometry of the type Ia SN 1999ee and the type Ib/c SN 1999ex in IC 5179. *Astron. J.* **124**, 2100–2117 (2002).
26. Hamuy, M. *et al.* The morphology of type Ia supernovae light curves. *Astron. J.* **112**, 2438–2447 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements All the co-authors participated in gathering the observations of the supernova. M.H. noticed the presence of hydrogen emission. M.M.P. noticed spectroscopic and photometric peculiarities and put forward the idea that these could be understood as due to SN/CSM interaction. N.B.S. provided the arguments about the progenitor types. M.H., M.M.P. and N.B.S. co-wrote this Letter. M.H. is a Hubble Fellow.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H. (mhamuy@ociw.edu).

Ballistic carbon nanotube field-effect transistors

Ali Javey¹, Jing Guo², Qian Wang¹, Mark Lundstrom² & Hongjie Dai¹

¹Department of Chemistry, Stanford University, California 94305, USA

²School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA

A common feature of the single-walled carbon-nanotube field-effect transistors fabricated to date has been the presence of a Schottky barrier at the nanotube–metal junctions^{1–3}. These energy barriers severely limit transistor conductance in the ‘ON’ state, and reduce the current delivery capability—a key determinant of device performance. Here we show that contacting semiconducting single-walled nanotubes by palladium, a noble metal with high work function and good wetting interactions with nanotubes, greatly reduces or eliminates the barriers for transport through the valence band of nanotubes. *In situ* modification of the electrode work function by hydrogen is carried out to shed light on the nature of the contacts. With Pd contacts, the ‘ON’ states of semiconducting nanotubes can behave like ohmically contacted ballistic metallic tubes, exhibiting room-temperature conductance near the ballistic transport limit of $4e^2/h$ (refs 4–6), high current-carrying capability ($\sim 25 \mu\text{A}$ per tube), and Fabry–Perot interferences⁵ at low temperatures. Under high voltage operation, the current saturation appears to be set by backscattering of the charge carriers by optical phonons. High-performance ballistic nanotube field-effect transistors with zero or slightly negative Schottky barriers are thus realized.

Transparent electrical contacts made to metallic single-walled carbon nanotubes (SWNTs) have revealed them to be ballistic conductors that exhibit two units of quantum conductance $4e^2/h$ ($R_Q = h/4e^2 = 6.5 \text{ k}\Omega$)^{4–6}. Carrier transport through the valence and conduction bands of a high-quality semiconducting SWNT could also be ballistic, presenting an opportunity to realize ballistic field-effect transistors (FETs) based on molecular electronic materials. However, such effort has been hampered by non-ideal electrical contacts made to semiconducting SWNTs. Previously, thermionic emission current and associated barriers between Ni contacts and SWNTs⁷ have been reported for materials grown by chemical vapour deposition (CVD). Recently, significant Schottky barriers (SBs) have been found in SWNT-FETs (with Ti contacts and laser-oven nanotubes)^{1–3}. Thermionic emission and tunnelling are involved in transport across the SBs, which limits the ON state

conductance of nanotube FETs to be well below the $4e^2/h$ limit ($G_{\text{ON}} \approx 0.001 \times 4e^2/h$ for laser tubes^{1–3}; and $G_{\text{ON}} \approx (0.05 - 0.3) \times 4e^2/h$ for CVD tubes^{7–10}) at room temperature. At low temperatures, quenching of thermionic currents causes semiconducting SWNT devices to be about 10–100 times more resistive than at room temperature^{7,11,12}.

A solution is presented here to eliminate or greatly suppress SBs at metal–nanotube contacts. We find that for Pd-contacted, long, semiconducting SWNT devices (length $L = 3 \mu\text{m}$, diameter $d \approx 3 \text{ nm}$, Fig. 1a) with back gates (SiO_2 thickness $t_{\text{ox}} = 500 \text{ nm}$), the room-temperature ON state conductance through the valence band (VB) is up to $G_{\text{ON}} \approx 0.1 \times (4e^2/h)$ ($R_{\text{ON}} = 60 \text{ k}\Omega$, Fig. 1b). For short channel nanotubes at room temperature ($L = 300 \text{ nm}$, Fig. 1a), $G_{\text{ON}} = (0.4 - 0.5) \times (4e^2/h)$ ($R_{\text{ON}} = 13 \text{ k}\Omega$, Fig. 1c). For long SWNTs, the ON state p-channel conductance exhibits metallic behaviour between room temperature and $\sim 200 \text{ K}$, and typically shows a downturn in G_{ON} upon further cooling (Fig. 1b inset). The short nanotubes do not show such downturn, and the average G_{ON} monotonically increases as temperature T decreases to $\sim 50 \text{ K}$ (Fig. 1d), below which pronounced oscillations with Fabry–Perot type of interferences⁵ appear in the G versus gate voltage (V_{gs}) data

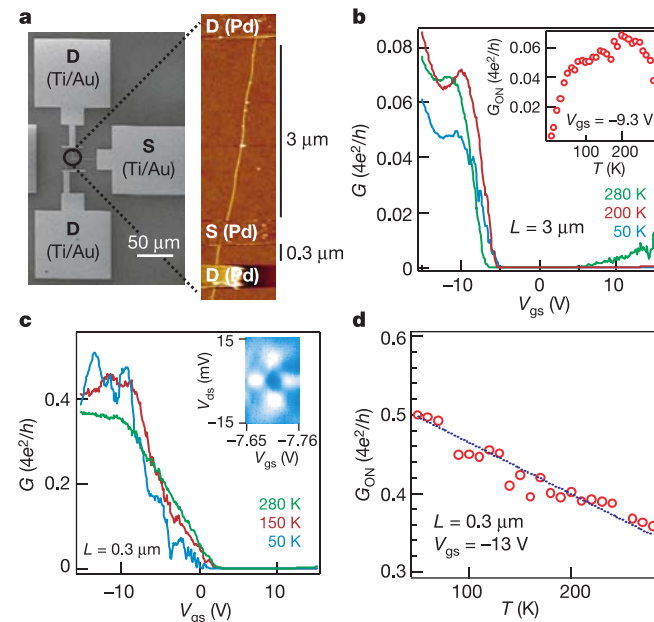


Figure 1 Pd-contacted long ($L = 3 \mu\text{m}$) and short ($L = 300 \text{ nm}$) back-gated SWNT devices formed on the same nanotubes on SiO_2/Si . **a**, A scanning electron microscope (SEM) image (left) and atomic force microscope (AFM) image (right) of a representative device. CVD synthesis for SWNTs and device fabrication were as described previously^{25,26}, except that Pd was used to contact nanotubes. The catalyst used here gave a wide range of nanotube diameters ($1.2 - 5 \text{ nm}$)²⁵. Ti/Au metal bonding pads were used to connect to the Pd source (S) and drain (D) electrodes. (We note that Pd electrodes tended to be soft and not robust against electrical probing). The devices were annealed in Ar at 225°C for 10 min after fabrication. The thickness of SiO_2 gate dielectric was $t_{\text{ox}} = 500 \text{ nm}$, except for the devices in Fig. 4 with $t_{\text{ox}} = 67 \text{ nm}$. AFM topographic height measurements were used to determine the diameters of SWNTs. The electrical data shown here were recorded with the devices placed in vacuum. **b**, G (at low S–D bias V_{ds}) versus gate voltage V_{gs} for a 3-μm-long SWNT ($d \approx 3.3 \text{ nm}$) device recorded at various T . Inset, G_{ON} versus T for the device. **c**, G versus V_{gs} for a 300-nm-long tube section on the same tube as for **b** at various T . Differential conductance $dI_{\text{ds}}/dV_{\text{ds}}$ versus V_{ds} and V_{gs} (inset, measured by a lock-in technique) at $T = 1.5 \text{ K}$ shows a Fabry–Perot-like interference pattern (bright peak $G \approx 4e^2/h$, dark region $G = 0.5 \times 4e^2/h$). Note that in certain V_{gs} regions, the pattern appears irregular, a phenomenon also seen for Fabry–Perot interference in metallic tubes⁵. **d**, G_{ON} versus T for the $L = 300 \text{ nm}$ semiconducting tube down to $\sim 50 \text{ K}$.

(peak conductance $\sim 4e^2/h$) (Fig. 1c inset). This corresponds to ballistic transport in the p-channel of semiconducting SWNTs. The metallic behaviour of the ON states of semiconducting SWNTs and G_{ON} approaching $4e^2/h$ strongly suggest nearly SB-free contacts made to nanotubes.

Because SBs are sensitive to the contact metal work function ϕ (ref. 1), we have carried out *in situ* modification of ϕ_{Pd} by utilizing the phenomenon that exposure of Pd to molecular hydrogen reduces its work function at room temperature¹³. Figure 2a shows that a SWNT-FET exhibits decreased p-channel conductance and increased n-channel conductance when exposed to H_2 . This result is consistent with increased SB height for hole transport, and lowered barrier height for electron transport, due to reduced ϕ_{Pd} (ref. 1) caused by dissolved hydrogen¹³. The original pure Pd-contacted SWNT-FETs exhibit metallic behaviour for p-channel conduction (Fig. 2c), and highly linear conductance at low V_{ds} (Fig. 2d). After

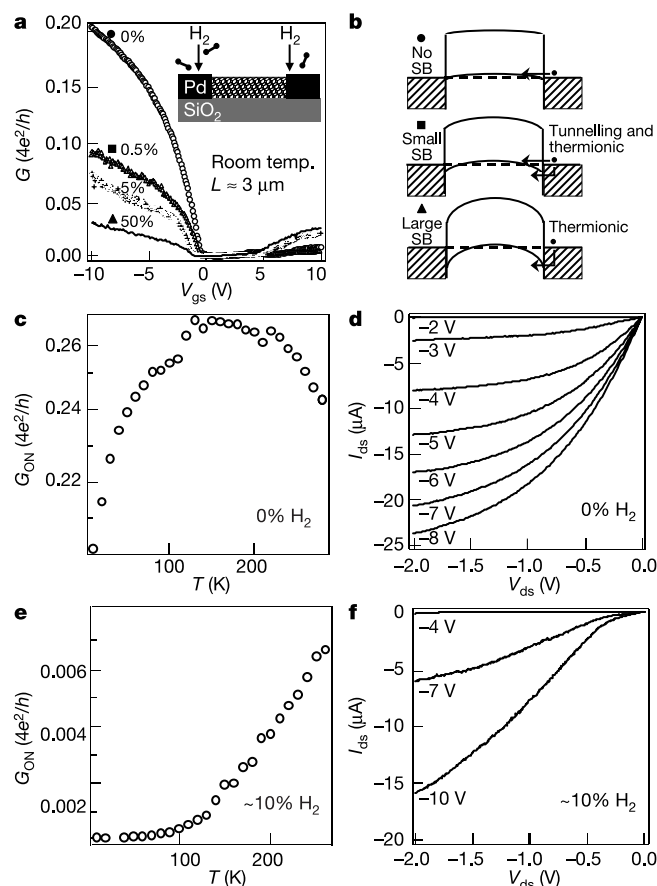


Figure 2 Properties of the metal/semiconducting-SWNT contacts as influenced by *in situ* metal work-function modification. **a**, Linear $G - V_{\text{gs}}$ for a SWNT-FET ($L = 3 \mu\text{m}$) before and after the pure Pd contacts were exposed to various concentrations of H_2 . Note that no change in any of the device characteristics was observed for Ti/Au- and Mo-contacted tube FETs when exposed to hydrogen. **b**, Schematic ON state (for p-channel) band diagrams for the device under various H_2 concentrations, depicting the development of SBs to the valence band for higher $[\text{H}_2]$ and lower ϕ_{Pd} (caused by dissolution of H_2 in Pd). The symbols in the diagrams (solid circle, square and triangle) correspond to those used in **a** to label the ON state gate voltages. **c**, G_{ON} versus T for the device with pure Pd contacts before exposure to H_2 . **d**, $I_{\text{ds}} - V_{\text{ds}}$ curves for the device in **c** under various V_{gs} at room temperature. **e**, G_{ON} versus T for the same device as in **c** and **d** after exposure of the Pd electrodes to $\sim 10\%$ H_2 . Both the p- and n- channel (data not shown) exhibit non-metallic behaviour, indicating mid-gap SBs to both valence and conduction bands with lowered ϕ_{Pd} by H_2 . **f**, $I_{\text{ds}} - V_{\text{ds}}$ curves for the device in **e** under various V_{gs} at room temperature exhibiting typical SB-FET behaviour with inflection points in the curves. The plots in **d** and **f** clearly show the presence and absence of a SB in the two cases (with and without H_2 exposure) respectively.

hydrogen modification of the Pd contacts, the p-channel of the device exhibits non-metallic behaviour (Fig. 2e), and inflection points (typical of SB-FETs) are observed in the $I_{\text{ds}} - V_{\text{ds}}$ characteristic of the device (Fig. 2f). These results suggest that higher SBs are formed to the p-channel of nanotubes with lower ϕ_{Pd} (Fig. 2b). The strong dependence of SB height on ϕ_{Pd} indicates the absence of appreciable interface states between SWNTs and Pd, and little or no Fermi level pinning¹⁴. We find that high- ϕ metal is not sufficient for forming transparent contacts to SWNT VBs. When experimenting with Pt with a high ϕ of 5.7 eV ($> \phi_{\text{Pd}} = 5.1$ eV), we find that Pt-contacted devices exhibit lower p-channel conductance and non-metallic behaviour for the ON states. This may be attributed to the poor sticking or wetting interaction between Pt and SWNTs, opposite to Pd (refs 15, 16). High- ϕ metal and high-quality metal-tube interfaces are thus both important for low contact barriers to the VBs of semiconducting nanotubes.

The transport properties of Pd-contacted SWNTs depend on nanotube diameter and length. For Pd-contacted SWNTs with $d \approx 1.7$ nm and $L \approx 275$ nm, $R_{\text{ON}} \approx 32$ k Ω can be obtained (Fig. 3a inset). For SWNTs with $d < \sim 2$ nm, SBs at the Pd contacts are significantly low, but their elimination is more difficult (bandgaps of SWNTs $\propto 1/d$). Reducing the nanotube channel length from 3 μm to 300 nm always lowers R_{ON} by up to about 20 times, a trend observed for various diameter SWNTs (Fig. 3). Transport in

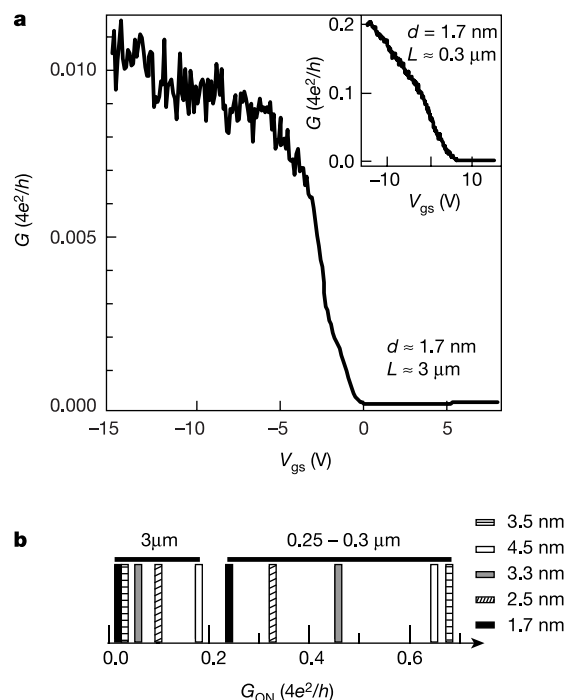


Figure 3 Geometry/transport-property relations for Pd-contacted semiconducting nanotubes. **a**, G versus V_{gs} for a $d \approx 1.7$ nm SWNT ($L = 3 \mu\text{m}$) contacted by Pd at room temperature. Inset, data for the same tube with $L = 275$ nm. **b**, G_{ON} for several $d \approx 1.5$ to 4.5 nm semiconducting SWNTs with lengths of ~ 300 nm and $\sim 3 \mu\text{m}$. Note that for our tube FETs with $t_{\text{ox}} = 500$ nm SiO_2 gate insulators, the threshold voltages (V_{T}) for the short devices tend to display a large shift (~ 5 V) to the more positive V_{gs} side from those of corresponding long devices (see **a** main panel versus **a** inset, and Fig. 1b versus Fig. 1c). This is attributed to the fact that the screening length in the quasi one-dimensional semiconducting nanotube devices is of the order of gate-dielectric thickness^{27–29}. For short channels, this can lead to effectively p-doping of the nanotube ($L \approx 300$ nm $< t_{\text{ox}}$) and thus the V_{T} shift. Scaling of the gate dielectric for the current devices is essential to scale the screening length and to minimize the variation of V_{T} with channel length. Indeed, the short (~ 300 nm) and long ($3 \mu\text{m}$) Pd-contacted tube FETs with $t_{\text{ox}} = 67$ nm show only small V_{T} shift (< 1 V, see Fig. 4a versus Fig. 4c, reproduced with over 30 such devices).

$L \approx 300$ nm tubes (of all diameters) is close to ballistic (R_{ON} down to $1.5 \times h/4e^2 \approx 10$ k Ω at room temperature). Acoustic phonons (twistons) could account for the extra resistance¹⁷, and quenching of these phonon modes at low T leads to G_{ON} approaching $4e^2/h$ and the observed metallic state, with clear manifestation of phase coherent resonances⁶. The higher R_{ON} for longer tubes suggests additional scattering along the length of long nanotubes. However, the precise nature of the scatters on the long semiconducting SWNTs is elusive, and could be due to structural defects, mechanical bending or chemical inhomogeneity. For long tubes, the upturn in the ON state resistance at low T is attributed to weak localization due to these defects, causing the deviation from metallic behaviour^{17,18}.

Figure 4 shows the detailed transistor data for long- and short-tube devices respectively with Pd contacts and back gates ($t_{\text{ox}} = 67$ nm for these FETs). Both devices exhibit $I_{\text{ON}}/I_{\text{OFF}} \approx 10^6$ (Fig. 4). The maximum 'on' state current reaches $I_{\text{Dsat}} \approx 25$ μA (current saturation at this level has been observed with metallic tubes previously¹⁹), about 4 times higher than achieved previously for CVD semiconducting tubes⁹ (highest $I_{\text{ON}} \approx 4$ μA for SB tube FETs^{12,20}). The current normalized by the width of the tube reaches $\sim 7,000$ $\mu\text{A} \mu\text{m}^{-1}$. A key goal in transistor scaling is to maximize the

saturation drain current for short channel and low voltage devices^{21,22}, and SB-free SWNT-FETs could be a promising approach to reach this goal. Negative SB-FETs will further enhance current delivery capability²¹, and may be possible with nanotubes due to the lack of significant Fermi level pinning at the contacts (based on the current work and ref. 14).

We now turn to the issue of how the Pd-contacted nanotube transistors operate and compare them with SB-FETs¹⁻³ (operation involves gate modulation of the SB width) and metal-oxide-semiconductor FETs (MOSFETs^{21,23}, operation involves gate modulation of carrier density in the channel). For the ON states of our SWNT-FETs at room temperature, the short tubes can be considered as nearly ballistic channels ($R_{\text{ON}} = 10$ k Ω) while the long tubes are diffusive channels ($R_{\text{ON}} = 37$ k Ω). For the OFF states, the large $I_{\text{ON}}/I_{\text{OFF}} \approx 10^6$ suggests that large thermal activation barriers exist to limit the OFF state currents. In between the ON and OFF states, the subthreshold swing is $S \approx 150$ mV per decade for the long tube and $S \approx 170$ mV per decade for the short tube. The value of S is a sensitive function of different device geometry parameters, but the value observed here falls closer to the S versus ϵ/t_{ox} (ϵ is the dielectric constant) curve for MOSFETs than that for SB-FETs

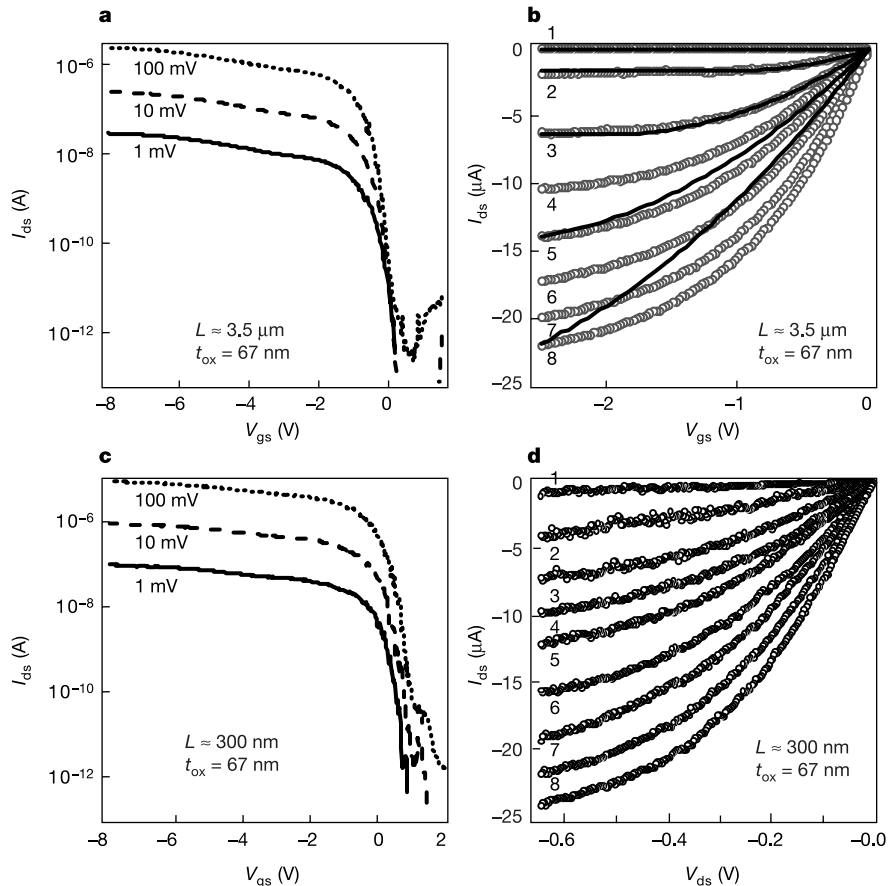


Figure 4 Room-temperature electrical properties of high-performance SWNT-FETs ($t_{\text{ox}} = 67$ nm). **a**, Transfer characteristics for a $d = 3.5$ nm and channel length ~ 3 μm SWNT with pure Pd contacts. $G_{\text{ON}} = 0.17 \times 4e^2/h$ ($R_{\text{ON}} \approx 37$ k Ω) under $V_{\text{ds}} = 1$ mV and $V_{\text{gs}} = -8$ V. **b**, $I_{\text{ds}} - V_{\text{ds}}$ versus V_{gs} for device in **a**. The curves (with circular symbols) labelled 1 to 8 correspond to $V_{\text{gs}} = -0.5$ to -7.5 V in steps of 1 V. The solid lines are calculated from the square-law model for MOSFETs to fit the experimental curves 1–5. **c**, Transfer characteristics for a device with channel length $L = 300$ nm formed on the same tube as in **a** and **b**. $G_{\text{ON}} = 0.65 \times 4e^2/h$ ($R_{\text{ON}} \approx 10$ k Ω) under $V_{\text{ds}} = 1$ mV and $V_{\text{gs}} = -8$ V. The current measured at $V_{\text{ds}} = 100$ mV for the OFF state is ~ 1 pA, as the hole current is limited by a body barrier in the valence band. The electron current,

limited by a SB to the conduction band, must be less than 1 pA, and to limit electron current to such a small value, a SB height of ~ 0.35 eV is needed by simple analysis. Since the sum of the electron- and hole-barrier heights is the bandgap (~ 0.3 eV for this tube), a slight negative SB for the holes is found to be about -0.05 V. **d**, $I_{\text{ds}} - V_{\text{ds}}$ versus V_{gs} characteristics for the short tube device in **c**. The curves labelled 1 to 9 correspond to 0.5, 0, -0.5 , -1 , -1.5 , -2.5 , -3.5 , -4.5 and -5.5 V, respectively. Note that the devices shown above were passivated by PMMA to eliminate hysteresis upon sweeping V_{gs} back and forth³⁰. This allowed for more precise determinations of V_{T} . The devices in Fig. 3 were also passivated by PMMA.

(~ 800 mV per decade predicted for the SB case²). These results plus the high ON state conductance (G_{ON} up to $0.65 \times 4e^2/h$ at room temperature) strongly suggest SB-free MOSFET operation of nanotube transistors.

We have analysed the 3- μm -long tube FETs by using the MOSFET square-law model for diffusive transport²³ (the gate capacitance was obtained by solving the two-dimensional Poisson equation). The output characteristics can be well explained by the model for $V_{\text{gs}} \geq -2.5$ V (Fig. 4b with fit data in solid line) and a hole mobility of $\mu_{\text{h}} \approx 4,000 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ is deduced. However, when the device is in highly ON states under $V_{\text{gs}} < -2.5$ V, we observe significant deviation of its characteristics from the square-law model (Fig. 4b) in the high-current regions of the $I_{\text{ds}} - V_{\text{ds}}$ curves. The saturation current I_{Dsat} (when $I_{\text{Dsat}} > 10 \mu\text{A}$) grows slower than expected from the square-law model as more negative V_{gs} is applied, especially when I_{Dsat} approaches $\sim 25 \mu\text{A}$ (Fig. 4b). Similar behaviour is well known for silicon MOSFETs when velocity saturation due to optical phonon scattering of high-energy carriers occurs, and the device changes from a square-law to linear dependence with gate voltage²³. In light of the earlier observation that in metallic tubes, the current starts to saturate at $\sim 10 \mu\text{A}$ and fully saturates at $\sim 25 \mu\text{A}$ due to optical or zone-boundary phonon scattering¹⁹, we suggest that a similar phenomena may be occurring at high currents ($> \sim 10 \mu\text{A}$) in our nanotube FETs. That is, high current flows in 'turned-on' semiconducting nanotubes are also limited by optical phonon backscattering (this effect has not been observed in earlier tube FETs with maximum currents at $\sim 7 \mu\text{A}$; see, for example, refs 9 and 10). We expect that nanotube channel length scaling below the optical phonon scattering mean free path ($\sim 100 \text{ nm}$)¹⁹ could lead to ballistic transport in the high-current regime, and further enhance the current delivery capability of nanotube FETs.

We have also analysed our short tube FETs by the square-law model, and found that the diffusive transport model completely fails to reproduce the device characteristics. This is an expected result, because transport in the short nanotubes is close to ballistic. Preliminary analysis based on a ballistic MOSFET model^{10,24} leads to good but non-ideal fitting of the experimental data. Further theoretical work is needed to model the detailed behaviour of nanotube transistors near the ballistic transport regime. $\square$

Received 19 April; accepted 2 June 2003; doi:10.1038/nature01797.

- Heinze, S. *et al.* Carbon nanotubes as Schottky barrier transistors. *Phys. Rev. Lett.* **89**, 106801 (2002).
- Appenzeller, J. *et al.* Field-modulated carrier transport in carbon nanotube transistors. *Phys. Rev. Lett.* **89**, 126801 (2002).
- Heinze, S., Radosavljevic, M., Tersoff, J. & Avouris, P. Unexpected scaling of the performance of carbon nanotube transistors. Preprint at (<http://xxx.lanl.gov/cond-mat/0302175>) (2003).
- White, C. T. & Todorov, T. N. Carbon nanotubes as long ballistic conductors. *Nature* **393**, 240–242 (1998).
- Liang, W. *et al.* Fabry-Perot interference in a nanotube electron waveguide. *Nature* **411**, 665–669 (2001).
- Kong, J. *et al.* Quantum interference and ballistic transmission in nanotube electron waveguides. *Phys. Rev. Lett.* **87**, 106801 (2001).
- Zhou, C., Kong, J. & Dai, H. Electrical measurements of individual semiconducting single-walled nanotubes of various diameters. *Appl. Phys. Lett.* **76**, 1597 (1999).
- Rosenblatt, S. *et al.* High performance electrolyte gated carbon nanotube transistors. *Nano Lett.* **2**, 869–915 (2002).
- Javey, A. *et al.* High-k dielectrics for advanced carbon-nanotube transistors and logic gates. *Nature Mater.* **1**, 241–246 (2002).
- Guo, J. *et al.* Assessment of silicon MOS and carbon nanotube FET performance limits using a general theory of ballistic transistors. *IEEE Int. Electron Devices Meeting Tech. Dig.* 711–714 (December 2002).
- Javey, A., Shim, M. & Dai, H. J. Electrical properties and devices of large-diameter single-walled carbon nanotubes. *Appl. Phys. Lett.* **80**, 1064–1066 (2002).
- Martel, R. *et al.* Ambipolar electrical transport in semiconducting single-wall carbon nanotubes. *Phys. Rev. Lett.* **87**, 256805 (2001).
- Mandelis, A. & Christofides, C. *Physics, Chemistry and Technology of Solid State Gas Sensor Devices* (Wiley, New York, 1993).
- Leonard, F. & Tersoff, J. Role of Fermi-level pinning in nanotube Schottky diodes. *Phys. Rev. Lett.* **84**, 4693–4696 (2000).
- Zhang, Y., Franklin, N., Chen, R. & Dai, H. Metal coating on suspended carbon nanotubes and its implication to metal-tube interactions. *Chem. Phys. Lett.* **331**, 35–41 (2000).
- Zhang, Y. & Dai, H. Formation of metal nanowires on suspended single-walled carbon nanotubes. *Appl. Phys. Lett.* **77**, 3015–3017 (2000).
- Kane, C. L. *et al.* Temperature dependent resistivity of single wall carbon nanotubes. *Euro. Phys. Lett.* **6**, 683–688 (1998).
- Fischer, J. E. *et al.* Metallic resistivity in crystalline ropes of single-wall carbon nanotubes. *Phys. Rev. B* **55**, R4921–R4924 (1997).

- Yao, Z., Kane, C. L. & Dekker, C. High-field electrical transport in single-wall carbon nanotubes. *Phys. Rev. Lett.* **84**, 2941–2944 (2000).
- Wind, S. J., Appenzeller, J., Martel, R., Derycke, V. & Avouris, P. Vertical scaling of carbon nanotube field-effect transistors using top gate electrodes. *Appl. Phys. Lett.* **80**, 3817–3819 (2002).
- Guo, J. & Lundstrom, M. A Computational study of thin-body, double-gate, Schottky barrier MOSFETs. *IEEE Trans. Elec. Dev.* **49**, 1897–1902 (2002).
- Guo, J., Lundstrom, M. & Datta, S. Performance projections for ballistic carbon nanotube field-effect transistors. *Appl. Phys. Lett.* **80**, 3192–3194 (2002).
- Sze, S. M. *Physics of Semiconductor Devices* (Wiley, New York, 1981).
- Rahman, A., Guo, J., Datta, S. & Lundstrom, M. Theory of ballistic nanotransistors. *IEEE Trans. Electron Dev. and IEEE Trans. Nanotechnol.* (Joint Special Issue on Nanoelectronics) (in the press).
- Kong, J., Soh, H., Cassell, A., Quate, C. F. & Dai, H. Synthesis of individual single-walled carbon nanotubes on patterned silicon wafers. *Nature* **395**, 878–881 (1998).
- Soh, H. *et al.* Integrated nanotube circuits: Controlled growth and ohmic contacting of single-walled carbon nanotubes. *Appl. Phys. Lett.* **75**, 627–629 (1999).
- Odintsov, A. A. Schottky barriers in carbon nanotube heterojunctions. *Phys. Rev. Lett.* **85**, 150–153 (2000).
- Leonard, F. & Tersoff, J. Novel length scales in nanotube devices. *Phys. Rev. Lett.* **83**, 5174–5177 (1999).
- Nakanishi, T., Bachold, A. & Dekker, C. Transport through the interface between a semiconducting carbon nanotube and a metal electrode. *Phys. Rev. B* **66**, 073307 (2002).
- Kim, W. *et al.* Hysteresis caused by water molecules in carbon nanotube field effect transistors. *Nano Lett.* **3**, 193–198 (2003).

Acknowledgements We thank P. McEuen, W. Harrison, D. Antoniadis, J. Bokor and J. Tersoff for insights and sharing preprints/unpublished results, and D. Wang for SEM. The work at Stanford was supported by the MARCO MSD Focus Center, DARPA/Molectronics, SRC/AMD, and the Packard, Alfred Sloan and Dreyfus foundations. The work at Purdue was supported by the National Science Foundation Nanoscale Modeling and Simulation Program and by the Network for Computational Nanotechnology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.D. (hdai@stanford.edu).

Low-loss hollow-core silica/air photonic bandgap fibre

Charlene M. Smith, Natesan Venkataraman, Michael T. Gallagher, Dirk Müller, James A. West, Nicholas F. Borrelli, Douglas C. Allan & Karl W. Koch

Corning Incorporated, Sullivan Park, Corning, New York 14831, USA

Photonic bandgap structures use the principle of interference to reflect radiation. Reflection from photonic bandgap structures has been demonstrated in one, two and three dimensions and various applications have been proposed^{1–4}. Early work in hollow-core photonic bandgap fibre technology⁵ used a hexagonal structure surrounding the air core; this fibre was the first demonstration of light guided inside an air core of a photonic bandgap fibre. The potential benefits of guiding light in air derive from lower Rayleigh scattering, lower nonlinearity and lower transmission loss compared to conventional waveguides. In addition, these fibres offer a new platform for studying nonlinear optics in gases⁶. Owing largely to challenges in fabrication, the early air-core fibres were only available in short lengths, and so systematic studies of loss were not possible. More recently, longer lengths of fibre have become available^{7,8} with reported losses of $1,000 \text{ dB km}^{-1}$. We report here the fabrication and characterization of long lengths of low attenuation photonic bandgap fibre. Attenuation of less than 30 dB km^{-1} over a wide transmission window is observed with minimum loss of 13 dB km^{-1} at $1,500 \text{ nm}$, measured on 100 m of fibre. Coupling between surface and core modes of the structure is identified as an important contributor to transmission loss in hollow-core photonic bandgap fibres.

We use a stack and draw method for the fabrication of photonic bandgap fibre (PBGF)^{5,7}. In this process, a preform consisting of stacked capillaries is built and then reduced to an intermediate-

sized cane, which is then reduced to fibre. We have chosen to focus on producing high air-filling fraction fibres; in order to achieve that goal the cane itself must have a high air-filling fraction.

The stack and draw method has been reported to produce 50 m of PBGFs that transmit in the infrared with attenuations^{5,7,8} of $1,000 \text{ dB km}^{-1}$. Here the stack and draw method is used to prepare a fibre preform consisting of silica glass and air, which is drawn into several hundred metres of low-loss PBGF. Over the course of the draw, draw stability was maintained by monitoring outside fibre diameter, which was held to $125 \pm 2 \mu\text{m}$.

A scanning electron microscopy (SEM) image of the fibre cross-section is shown in Fig. 1. The image is cropped to focus on the air core and first few rings of microstructured cladding. The fibre contains 8 rings of air holes surrounding the core. At each end of a 100-m-long fibre we measure the fibre structure and find that the lattice spacing (pitch), $\Lambda = 4.7 \mu\text{m}$, an air-filling fraction of 0.94, and core diameter $D = 12.7 \mu\text{m}$ remain approximately constant over the 100-m fibre length. The core diameter is estimated as a circle whose area is equivalent to the area enclosed by the central defect.

We note that the air-filling fraction of 0.94 cannot be obtained by packing non-intersecting circular air holes. Inspection of the SEM image reveals that the holes are better represented by rounded hexagons; this geometry can achieve higher air-filling fractions than circles. Air-filling fraction is critical to determining key attributes of PBGF, such as the width and position of bandgap.

We measure the spectral transmission and modal properties of a $\sim 100\text{-m}$ -long fibre using a broadband edge-emitting light-emitting diode (EELED) source with a wavelength range from 1,200 to 1,700 nm. Light from a fibre-coupled EELED source is butt-coupled via an SMF-28 optical fibre into the PBGF. We verify that the light is guided within the air core by imaging the output facet of the fibre with an infrared camera. The mode images (Fig. 2) show the strong confinement of these modes within the photonic bandgap structure.

With the fundamental mode excited, we measure the transmission spectrum of the fibre. For this measurement we collect the output light with a multimode fibre and feed the signal into an optical spectrum analyser. The normalized transmission spectrum of the entire 100-m length is shown in Fig. 3. We measure the transmission window spanning 1,395 to 1,700 nm, and observe high attenuation in the region from 1,550 to 1,650 nm.

We also measure the transmission spectra of $\sim 1.5\text{-m}$ -long pieces from both ends of the original 100-m-long fibre. Comparison of the spectral characteristics of the 1.5-m-long pieces taken from the fibre beginning, $L = 0 \text{ m}$, and fibre end, $L = 100 \text{ m}$, with the full 100 m length, shows that the transmission window shifts only slightly over the full fibre length. Taking the 3-dB transmission point as a comparator, the leading edge of the transmission window shifts from 1,358 to 1,388 nm. Inasmuch as the width of the transmission window of a photonic crystal is governed by its geometric parameters (pitch, hole size, core size) the similarity of the overall transmission spectrum to the spectra of the beginning and end pieces suggests that uniformity along the 100 m is maintained. For

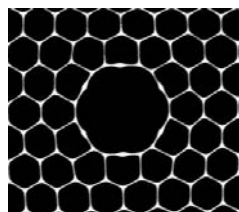


Figure 1 SEM image of air-core PBGF profile, truncated to show central core defect and first rings of holes of microstructured cladding. The pitch of the holes is roughly $4.7 \mu\text{m}$ and the air-filling fraction is 0.94. The central air hole has an area equivalent to a circle of $12.7 \mu\text{m}$ diameter.

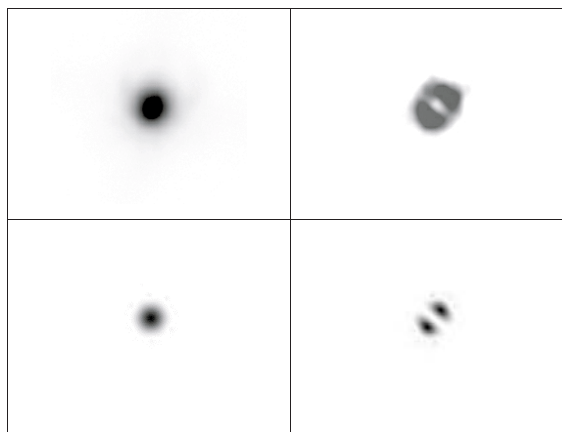


Figure 2 Comparison of experimental and theoretical mode profiles for the air-core PBGF. Top, experimental; bottom, theoretical; left, fundamental mode; and right, first higher-order mode. Experimentally, excitation of the fundamental or higher-order mode is accomplished by optimizing the angle of launch fibre.

example, if one assumes that only the scale of the lattice changes along the length of the fibre, the observed spectral shift of the leading edge of the transmission window corresponds to a change in scale of less than 4%. We attribute this high degree of structural uniformity and concomitant spectral integrity to precise control over draw parameters as illustrated by the fibre diameter measurements described previously.

The spectral attenuation of the 100-m section of fibre is shown in Fig. 4. The measured data show the lowest loss of 13 dB km^{-1} at a wavelength of 1,500 nm and a loss of the order of 200 dB km^{-1} within the more absorbing region.

We emphasize that the measured transmission window from 1,395 to 1,700 nm represents one bandgap; the spectral features in the higher loss region are not believed to be due to a bandgap edge, but rather due to coupling of the core mode to discrete lossy modes supported by the structure in this wavelength region. Calculations of the fibre structure support this proposed loss mechanism.

We use a publicly available vector plane-wave simulation⁹ of the Maxwell equations to calculate the modes of the real dielectric structure obtained from an SEM bitmap image similar to Fig. 1. (For a detailed discussion of photonic bandgaps and defect modes,

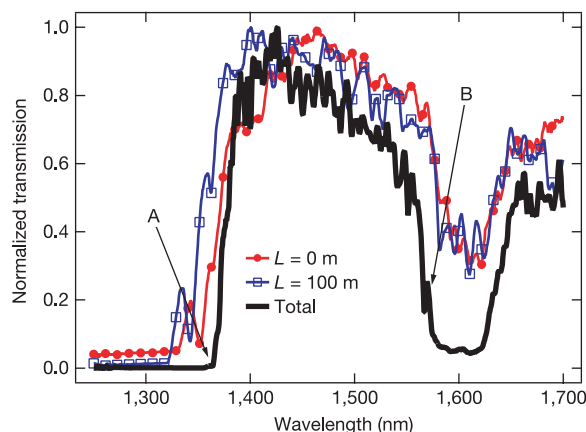


Figure 3 Transmission spectra for the full 100-m-long fibre and two 1.5-m-long pieces taken from the end and the beginning of the 100-m section. We measure air-core transmission between 1,395 and 1,700 nm. The comparison between the beginning and end pieces shows a less than 4% variation in the edges of the transmission windows. The sharp transmission edge (A) corresponds to the bandgap edge. The attenuation between 1,550 and 1,650 nm (B) is attributed to surface and core mode coupling.

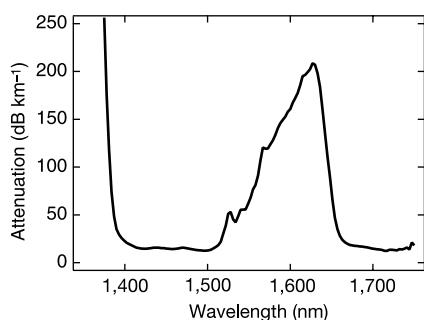


Figure 4 Optical attenuation as a function of wavelength for the 100-m-long air-core PBGF. A cutback method using a commercial Photon-Kinetics measurement bench was used. An intermediate SMF-28 launch fibre was inserted between the focusing lens of the measurement bench and the input end of the fibre under test. This arrangement allows for selective excitation of the air-core modes and minimizes light coupled into the microstructured fibre cladding, ensuring that we are guiding in the air core for our loss measurement. The attenuation of the PBGF is measured by subtracting the throughput of the full fibre from the throughput of a 2-m-long reference section.

see ref. 10.) Fully vectorial eigenmodes of Maxwell's equations with periodic boundary conditions were computed by preconditioned conjugate-gradient minimization of the block Rayleigh quotient in a plane-wave basis. Our simulation produces mode profiles in good agreement with the measured modes (Fig. 2). Owing to the exceptionally thin walls and high air-filling fraction of the fibre, the interpretation of the limited-resolution bitmap of the structure is critical to the prediction of the spectral position and width of the gap. The conversion used to produce a two-level image is justified by the excellent agreement between the model and experiment.

In Fig. 5 we plot the effective refractive index of the modes as a function of wavelength. The numerical modelling uses periodic boundary conditions, which, for the fibre profile bitmap, introduces modes at the boundary of the supercell where the edge of the bitmap does not mesh with its periodic image. We ignore these superfluous modes in the results shown in Fig. 5. The black symbols represent continuum modes that are adjacent to the bandgap region. We calculate a finite number of modes near the bandgap region, so the continuum modes are only shown near the top and bottom edge of the bandgap. In qualitative agreement with the measured data, the calculated bandgap spans the wavelength range between 1,400 and 1,850 nm. (Note that recent transmission measurements of this fibre using a long wavelength measurement bench show the long wavelength edge of the transmission window to be roughly 1,850 nm).

The simulation shows six air-core modes: two polarizations of a fundamental air-core mode and two polarizations of two, nearly degenerate, higher-order modes. Comparing the experimental transmission spectrum of the fundamental mode to the calculated data of this mode, we find good quantitative agreement. In addition to the air-core modes (Fig. 5 inset), there are four modes that cross the air-core modes within the bandgap. These modes are associated with the boundary of the air core. These surface modes are related to the silica structure around the core and reside near the silica/air interface. We note in the diagram the crossing or near degeneracy of the defect modes with the surface mode between 1,575 and 1,710 nm. We attribute the lossy transmission features from 1,550 to 1,700 nm in our measured PBGF to mode coupling between surface modes and core modes. The degeneracy and spatial overlap of the core modes with the surface modes provide an avenue by which the surface modes act as intermediaries to couple light out of the core and into the quasi-continuum of lossy cladding and radiation modes. A future publication will detail the quantitative effect of surface-to-core-mode coupling.

In summary, we have prepared an air-core photonic bandgap

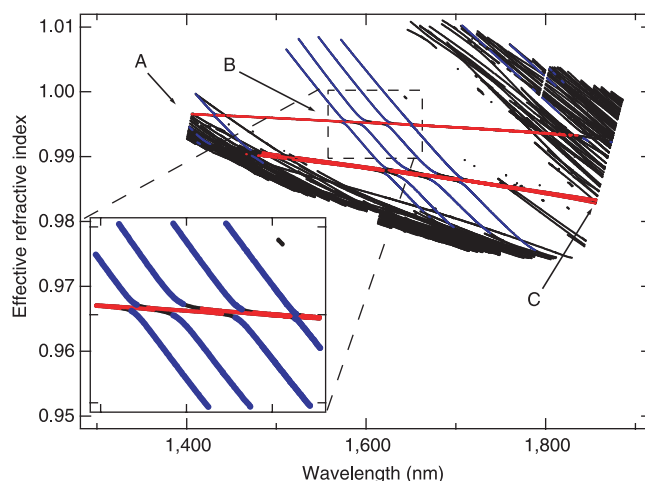


Figure 5 Calculated dispersion of air-core modes and bandgap modes for the fibre profile in Fig. 1. The black symbols represent modes of the structure. The red symbols represent the modes with a majority of their energy within the air-core defect. The fundamental air-core modes have a higher effective index than the higher-order air-core modes. The blue symbols represent modes whose energy is predominantly on the surface of the air-core defect. The locations A and C denote the edges of the photonic bandgap. The label B and the inset show the avoided crossings of surface modes with the fundamental core mode.

fibre having a transmission window greater than 400 nm with a minimum loss of 13 dB km^{-1} at 1,500 nm. Long lengths of this fibre were fabricated using a stack-and-draw technique. We also observed a higher-loss spectral region, which we associate with the interaction of the core modes with surface modes. Calculations of the bandgap using an SEM for the fibre cross-section suggest that the 400-nm transmission window represents one bandgap; the decreased transmission in a spectral area within the single bandgap is attributed to coupling of the core modes to surface modes. We propose that the surface modes, by coupling to the core as well as lossy continuum or radiation modes, provide a loss mechanism for the core mode. The reduction in loss is two orders of magnitude over previously reported loss results, demonstrating a significant advance towards low-loss air-core photonic bandgap fibre. Further, the results suggest the potential for this technology to achieve interaction lengths of several hundred metres, and may lead to devices that require two orders of magnitude less pump power. □

Received 24 February; accepted 18 June 2003; doi:10.1038/nature01849.

1. Krauss, T. F. & De La Rue, R. M. Photonic crystals in the optical regime—past, present and future. *Prog. Quantum Electron.* **23**, 51–96 (1999).
2. Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals. *Solid State Commun.* **102**, 165–173 (1997).
3. Krauss, T. F. & Baba, T. (eds) Photonic crystal structures and applications. *IEEE J. Quantum Electron.* **38**, 724–963 (2002).
4. Soukoulis, C. M. (ed.) *Photonic Crystals and Light Localization in the 21st Century* (Kluwer Academic, Dordrecht, 2001).
5. Cregan, R. F. *et al.* Single mode photonic band gap guidance of light in air. *Science* **285**, 1537–1539 (1999).
6. Benabid, F., Knight, J. C., Antonopoulos, G. & St Russell, P. J. Stimulated Raman scattering in hydrogen-filled hollow-core photonic crystal fibres. *Science* **298**, 399–402 (2002).
7. Allan, D. C. *et al.* in *Photonic Crystals and Light Localization in the 21st Century* (ed. Soukoulis, C. M.) 305–320 (Kluwer Academic, Dordrecht, 2001).
8. West, J. A. *et al.* Demonstration of an IR-optimized air-core photonic band-gap fibre. In *ECOC 2001* paper Th.A.2 (Eur. Conf. on Optical Communications, Amsterdam, 2001).
9. Johnson, S. G. & Joannopoulos, J. D. Block-iterative frequency-domain methods for Maxwell's equations in a plane-wave basis. *Opt. Express* **8**, 173–190 (2001).
10. Joannopoulos, J. D., Meade, R. & Winn, J. *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, Princeton, New Jersey, 1995).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.W.K. (kochkw@corning.com).

Unusually large earthquakes inferred from tsunami deposits along the Kuril trench

Futoshi Nanayama¹, Kenji Satake¹, Ryuta Furukawa¹, Koichi Shimokawa¹, Brian F. Atwater², Kiyoyuki Shigeno³ & Shigeru Yamaki⁴

¹Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8567, Japan

²US Geological Survey at University of Washington, Seattle, Washington 98195-1310, USA

³Meiji Consultant Co. Ltd, Chuo-ku, Sapporo 064-0807, Japan

⁴Seamus Ltd, Toyosaka, Niigata 950-3304, Japan

The Pacific plate converges with northeastern Eurasia at a rate of 8–9 m per century along the Kamchatka, Kuril and Japan trenches¹. Along the southern Kuril trench, which faces the Japanese island of Hokkaido, this fast subduction has recurrently generated earthquakes with magnitudes of up to ~8 over the past two centuries^{2–6}. These historical events, on rupture segments 100–200 km long, have been considered characteristic of Hokkaido's plate-boundary earthquakes^{7,8}. But here we use deposits of prehistoric tsunamis to infer the infrequent occurrence of larger earthquakes generated from longer ruptures. Many of these tsunami deposits form sheets of sand that extend kilometres inland from the deposits of historical tsunamis. Stratigraphic series of extensive sand sheets, intercalated with dated volcanic-ash layers, show that such unusually large tsunamis occurred about every 500 years on average over the past 2,000–7,000 years,

most recently ~350 years ago. Numerical simulations of these tsunamis are best explained by earthquakes that individually rupture multiple segments along the southern Kuril trench. We infer that such multi-segment earthquakes persistently recur among a larger number of single-segment events.

Eastern Hokkaido's largest well-documented interplate earthquake—the Tokachi-oki earthquake of moment magnitude (M_w) 8.1 in 1952 (ref. 4)—shook much of northeast Japan and generated tsunami waves 1–4 m high along the Hokkaido coast⁹. The ensuing 1973 Nemuro-oki earthquake (M_w 7.8)^{2,3}, from a shorter rupture to the east, had similar but more localized effects. Eastern Hokkaido previously suffered great earthquakes in 1843 (tsunami magnitude M_t 8.0) and 1894 (M_t 8.2)^{10,11}. The ruptures in 1843 and 1894 probably resembled those in 1952 and 1973, respectively^{5,6} (Fig. 1). Eastern Hokkaido may also have a late Holocene history of coseismic or postseismic emergence at estuaries¹². The region's most direct and compelling evidence for prehistoric earthquakes, however, consists of sheets of sand near the open coast.

Although tsunamis have left sand sheets on many coasts^{13,14}, few reported examples rival eastern Hokkaido's in landward extent or in number of successive sheets (Fig. 2, Supplementary Information,

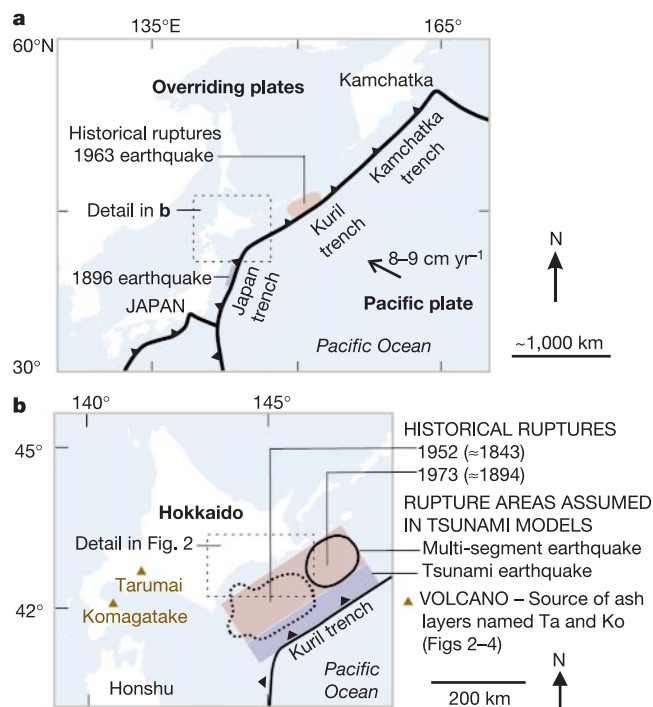


Figure 1 Maps of study area. Barbed lines show seaward edges of subduction zones; teeth point down the plate boundary. **a**, Subduction zone along the Kamchatka, Kuril and Japan trenches, showing rupture areas of 1896 and 1963 earthquakes^{20,23}. **b**, Hokkaido and northernmost Honshu, showing rupture areas of the 1952 and 1973 events^{2,4}, estimated locations of 1843 and 1894 earthquakes (rupture areas poorly known^{5,6}), and rupture areas of hypothetical earthquakes used in tsunami simulations (Fig. 2c).

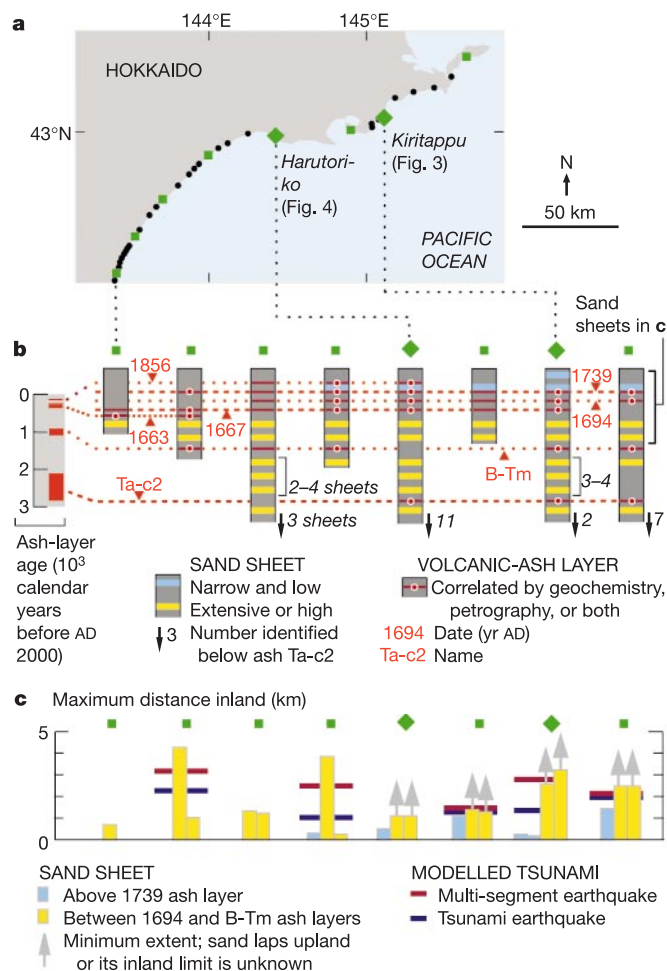


Figure 2 Summary of tsunami deposits in eastern Hokkaido and comparison with modelled inundation. Supporting details in Supplementary Tables S1 and S2, and Supplementary Fig. S1. **a**, Sites where prehistoric tsunami deposits have been found. **b**, Correlation of tsunami deposits by volcanic-ash layers. Samples were identified geochemically as in Supplementary Table S2, except for Usu-b, which was identified by petrography only. B-Tm, volcanic ash from Baitousan volcano, on the China–Korea border. **c**, Inland extent of sand sheets, compared with extent of inundation computed for tsunamis from hypothetical earthquakes in Fig. 1b.

and refs 15, 16). Many of Hokkaido's prehistoric sand sheets penetrate hundreds to thousands of metres inland (Fig. 2c). At all sites, the youngest extensive sheet pre-dates a widespread volcanic ash layer from 1694 (ref. 17) and in the southwest it also pre-dates localized ash layers from the 1660s (Fig. 2b). No tsunami or storm surge since 1694 has produced comparable deposits in the region¹⁸. The following examples show that outsize, prehistoric tsunamis repeatedly overran a beach-ridge plain (at Kiritappu; Fig. 3) and invaded a nearby coastal lake or lagoon (Harutori-ko; Fig. 4) at intervals averaging about 500 yr.

The sand sheets at Kiritappu are interbedded with freshwater peat and volcanic ash (Fig. 3). At least five sheets extend about 3 km across the beach-ridge plain, which averages a few metres above sea level. Marine diatoms abound in the sand but not in the intervening peat (Fig. 3d). Four or five extensive sand sheets overlie the volcanic ash Ta-c2 (2,000–3,000 yr old; Supplementary Table S2), and all the sheets that extend more than 200 m inland underlie the Ko-c2 ash (AD 1694; ref. 17).

The long stratigraphic record at Harutori-ko provides evidence for at least 17 tsunamis in the past 7,000 yr (Fig. 4). This lake or lagoon, which contains anoxic bottom water¹⁹, is flanked by Pleistocene uplands and separated from the sea by a 5-m-high beach berm that the 1952 tsunami did not cross (Fig. 4a, b). Deposits from the past 7,000 yr beneath the floor of Harutori-ko consist of laminated mud that alternates with sandy beds 0.1–1.0 m thick. The laminated mud, undisturbed by bottom-dwelling organisms, implies that a seaside berm has prevented tidal mixing in Harutori-ko through much of the Holocene. A core 1 km from the modern beach contains 17 of the sandy beds (Fig. 4c). Some of these beds grade upward from shell-bearing sand, through mud-clast breccia and laminated silt, into organic mud (Fig. 4d). At least four

of the beds lie between the 2,000- to 3,000-yr-old ash and the AD 1694 ash (Fig. 4c).

Large displacements of the sea floor are required to have generated the outsize waves implied by the eastern Hokkaido's prehistoric tsunami deposits. Because bathymetric mapping and seismic profiling along the Kuril and Japan trenches have yet to reveal enormous submarine landslides, we attempted to simulate the waves as products of tectonic displacement during two kinds of submarine earthquakes. One has a broad, multi-segment rupture 300 km by 100 km at 17–51 km depth, with 5 m slip. This M_w 8.4 earthquake, nearly as large as the nearby 1963 Kuril Islands earthquake²⁰ (Fig. 1a), simulates combined rupture of the Hokkaido segments that ruptured in the 1952 and 1973 earthquakes (Fig. 1b). The other is a tsunami earthquake—an event that produces water waves much larger than expected from seismic waves^{21,22}. We assume 5 m of slip on a narrow rupture 300 km by 50 km that extends to the sea floor directly updip from our multi-segment rupture (Fig. 5a). In size and depth, this hypothetical event resembles the tsunami earthquake along the Japan trench in 1896 (Fig. 1a; ref. 23). The 1896 tsunami reached heights over 10 m in northeast Honshu but less than 1 m in the eastern Hokkaido^{11,23}, and it may have been a repeat of an tsunami that devastated northeast Honshu in 1611 (ref. 11).

Although they predict similar tsunami heights at the coast, the two models differ in pattern of sea-floor displacement and in predicted inland extent of inundation. For the Pacific coast of eastern Hokkaido, both models predict tsunami heights mostly in the range 3–10 m (Supplementary Fig. S1b). However, for earthquakes having the same amount of slip, the sea-floor uplift in the multi-segment model is three times wider and 1 m higher than in the tsunami earthquake model (Fig. 5b). Consequently, only the tsunami waves from the broad, multi-segment rupture run far

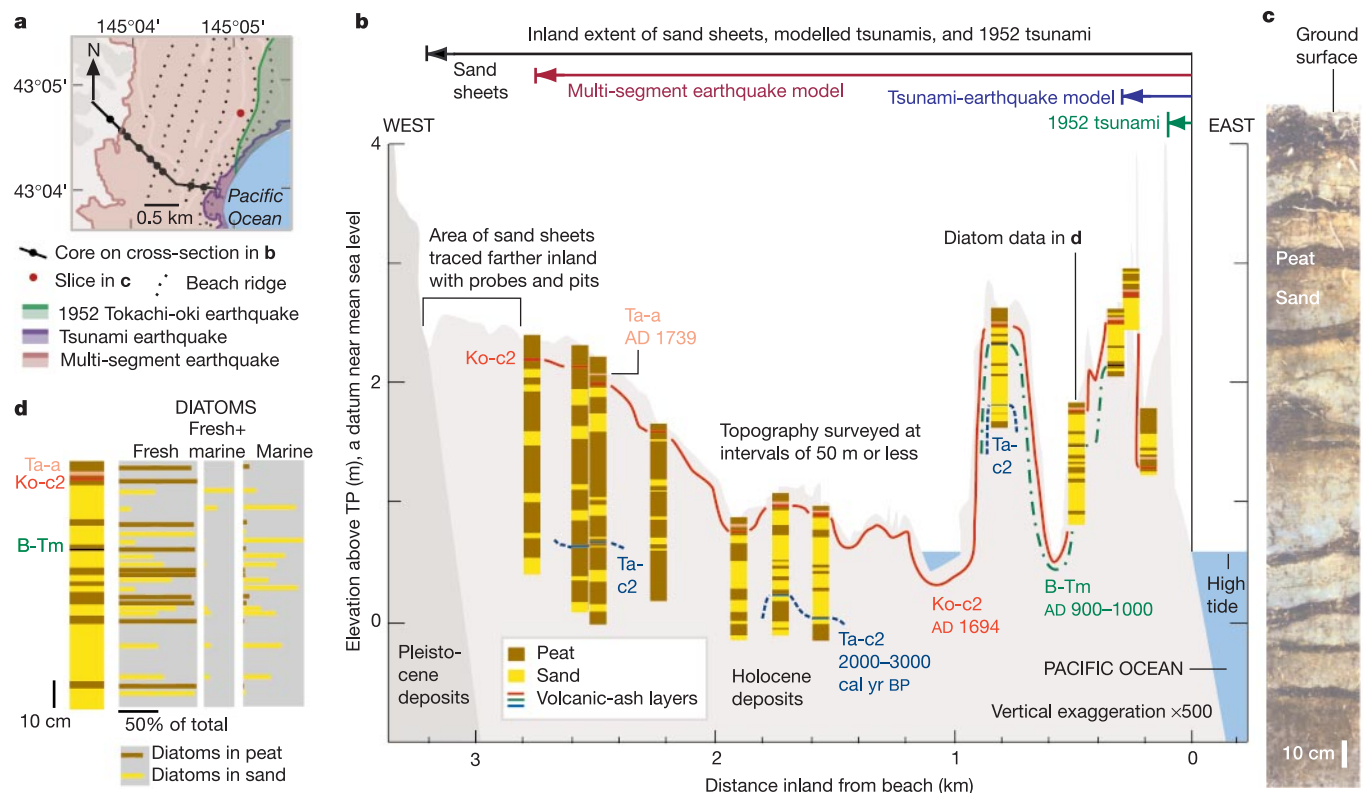


Figure 3 Evidence for outsize tsunamis at Kiritappu. **a**, Index map, showing inundation areas reported for the 1952 tsunami⁹ and computed for the two hypothetical earthquakes in Fig. 1b. **b**, Cross-section of cores correlated by volcanic ash from Tarumai (Ta-a, Ta-c2), Komagatake (Ko-c2) and Baitousan (B-Tm) (ref. 17, Supplementary Table S2). Arrows

at top compare sand-sheet extent with inundation computed from hypothetical earthquakes and with inundation observed for 1952 tsunami⁹. **c**, Sand interbedded with peat in slice collected 0.5 km from the coast (at red dot in **a**). **d**, Diatom assemblages from a core in **b** (supporting details in Supplementary Table S3).

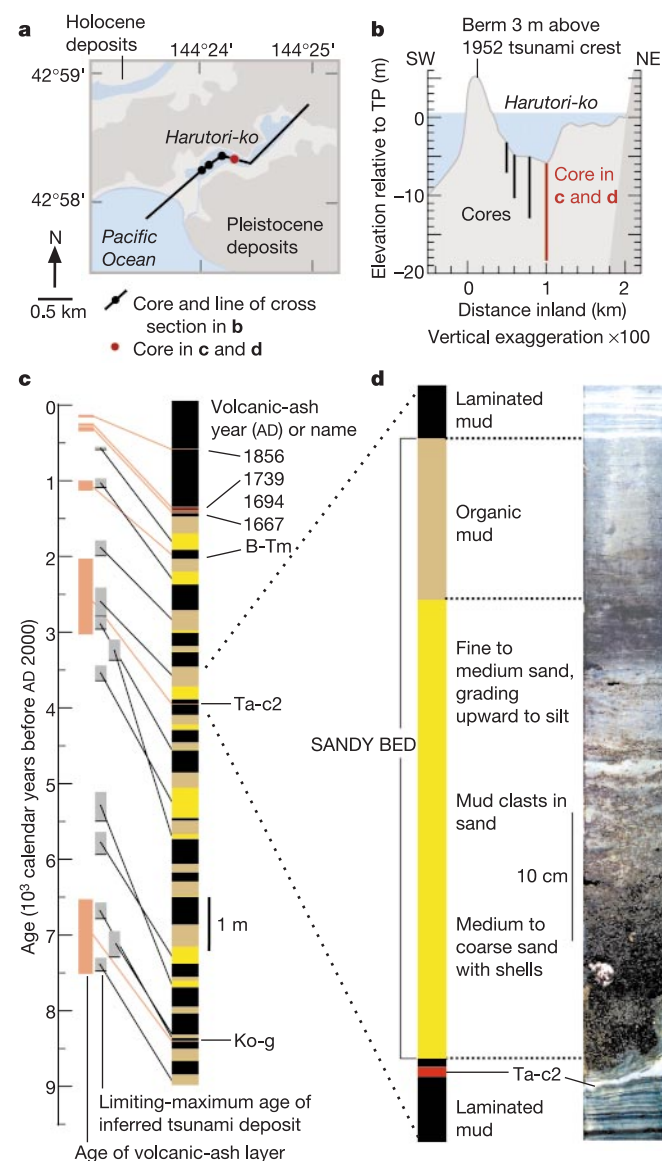


Figure 4 Evidence for outside tsunamis at Harutori-ko. **a**, **b**, Geomorphic setting. **c**, Generalized stratigraphy from the past 7,000 yr. **d**, Typical sandy bed.

enough inland to approach the sand-sheet limits (Figs 2c and 3a, b). Compared with inundation from the narrow tsunami-earthquake rupture, these waves penetrate farther because their wavelength is broader.

Tsunami deposits from recent millennia in eastern Hokkaido thus imply larger earthquakes and longer recurrence intervals than do records from the region's two historical centuries. The contrasting earthquakes are probably analogous to the successive twentieth-century ruptures that differed in length by factors of two or more at other subduction zones^{24–26}. Similarly, along the Nankai trough of southwest Japan, a multi-segment rupture 500 km long in 1707 broke again, in shorter pieces, as a pair of lesser earthquakes in 1854 and another pair in 1944–46 (ref. 26). Eastern Hokkaido's combination of prehistoric and historical evidence goes beyond these historical examples by implying persistent recurrence of multi-segment earthquakes at a subduction zone that is also subject to single-segment events. This persistence, in turn, implies that variation in rupture mode is characteristic of subduction along the southern Kuril trench.

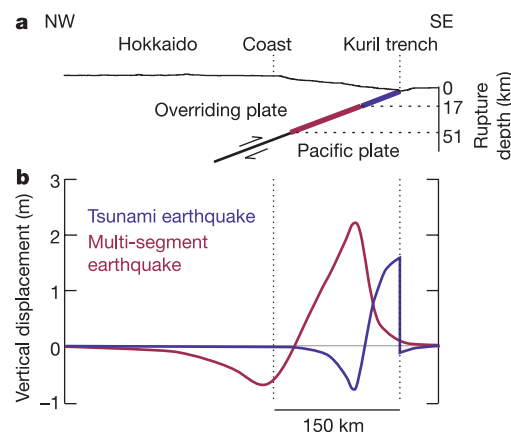


Figure 5 Cross-sectional view of **a**, two fault models and **b**, the sea-floor displacement computed from them.

Methods

In numerical simulations of tsunami inundation, we first computed sea-floor displacement from fault parameters (Fig. 5)²⁷, then used a finite-difference method to solve the nonlinear long-wave equation and the equation of continuity^{28,29}. The minimum grid interval is 225 m along the entire coast in Fig. 2a. To compute inland penetration at Kiritappu (Fig. 3a,b) and other sites in Fig. 2c, we used 25-m grids for bathymetry and topography, and we assumed the moving-boundary condition³⁰. In addition to the two tsunami sources described in the text, we also considered a tsunami earthquake with 10 m of slip (about twice that inferred for the 1896 Sanriku tsunami earthquake), which produces larger coastal heights than shown in Supplementary Fig. S2b but little more inundation distance than shown in Fig. 2c.

Received 20 January; accepted 24 June 2003; doi:10.1038/nature01864.

- DeMets, C. Oblique convergence and deformation along the Kuril and Japan Trenches. *J. Geophys. Res.* **97**, 17615–17625 (1992).
- Aida, I. Reliability of a tsunami source model derived from fault parameters. *J. Phys. Earth* **26**, 57–73 (1978).
- Shimazaki, K. Nemuro-oki earthquake of June 17, 1973: A lithospheric rebound at the upper half of the interface. *Phys. Earth Planet. Inter.* **9**, 314–327 (1974).
- Hirata, K., Geist, E. L., Satake, K., Tanioka, Y. & Yamaki, S. Slip distribution of the 1952 Tokachi-Oki earthquake (*M* 8.1) along the Kuril trench deduced from tsunami waveform inversion. *J. Geophys. Res.* **108**, 101029/2002JB001976 (2003).
- Hatori, T. Source area of the tsunami off the Nemuro Peninsula in 1973 and its comparison with the tsunami in 1894 [in Japanese with English abstract]. *Spec. Bull. Earthq. Res. Inst. Univ. Tokyo* **13**, 67–76 (1974).
- Hatori, T. Source area of the East Hokkaido Tsunami generated in April, 1843 [in Japanese with English abstract]. *Bull. Earthq. Res. Inst. Univ. Tokyo* **59**, 423–431 (1984).
- Fukao, Y. & Furumoto, M. Stress drops, wave spectra and recurrence intervals of great earthquakes — implications of the Etorofu earthquake of 1958 November 6. *Geophys. J. R. Astron. Soc.* **57**, 23–40 (1979).
- Nishenko, S. P. Circum-Pacific seismic potential: 1989–1999. *Pure Appl. Geophys.* **135**, 169–259 (1991).
- Central Meteorological Observatory The Tokachi earthquake of March 4, 1952 [in Japanese with English abstract]. *Q. J. Seismol.* **17**, 1–135 (1953).
- Abe, K. Quantification of historical tsunamis by the *M_s* scale [in Japanese with English abstract]. *Zisin* **52**, 369–377 (1999).
- Watanabe, H. *Comprehensive List of Tsunamis to Hit the Japanese Islands* [in Japanese] (Univ. Tokyo Press, Tokyo, 1998).
- Sawai, Y., Nasu, H. & Yasuda, Y. Fluctuations in relative sea-level during the past 3000 yr in the Onnetoh estuary, Hokkaido, northern Japan. *J. Quat. Sci.* **17**, 607–622 (2002).
- Clague, J. J., Bobrowsky, P. T. & Hutchinson, I. A review of geological records of large tsunamis at Vancouver Island, British Columbia, and implications for hazard. *Quat. Sci. Rev.* **19**, 849–863 (2000).
- Pinegina, T. K., Bourgeois, J., Bazanova, L. L., Melekestsev, I. V. & Braitseva, O. A. A millennial-scale record of Holocene tsunamis on the Lronotskiy Bay coast, Kamchatka, Russia. *Quat. Res.* **59**, 36–47 (2003).
- Hirakawa, K., Nakamura, Y. & Echigo, T. Tokachi chiho Taiheiyō engan chiiki no kyodai tsunami [Giant tsunami along the Pacific coast of the Tokachi region]. *Gekkan Chikyu* (supplement 31), 92–98 (2000).
- Sawai, Y. Evidence for 17th-century tsunamis generated on the Kuril-Kamchatka subduction zone, Lake Tokotan, Hokkaido, Japan. *J. Asian Earth Sci.* **20**, 903–911 (2002).
- Furukawa, R., Yoshimoto, M., Yamagata, K., Wada, K. & Ui, T. Did Hokkaido Komagatake volcano erupt in 1694? — reappraisal of the eruptive ages of 17–18th centuries in Hokkaido [in Japanese with English abstract]. *Kazan* **42**, 269–276 (1997).
- Nishimura, Y., Miyaji, N., Yoshida, M., Murata, T. & Nakagawa, M. The 1843 tsunami deposits found in the peat deposit at Kiritappu marsh, eastern Hokkaido, Japan [in Japanese with English abstract]. *Daiyōki Kenkyū* **39**, 451–460 (2000).
- Chikita, K., Fukuyama, R., Sakamoto, H. & Nakamichi, K. Dynamic behaviors of “dead water” in a coastal lagoon, Lake Harutori, Kushiro, Hokkaido: Field measurements for the ice-covered period [in Japanese with English abstract]. *Geophys. Bull. Hokkaido Univ.* **60** 13–28 (1997).

20. Beck, S. L. & Ruff, L. J. Rupture process of the great 1963 Kuril Islands earthquake sequence: asperity interaction and multiple event rupture. *J. Geophys. Res.* **92**, 14123–14138 (1987).
21. Kanamori, H. Mechanism of tsunami earthquakes. *Phys. Earth Planet. Inter.* **6**, 246–259 (1972).
22. Okal, E. A. & Newman, A. V. Tsunami earthquakes: The quest for a regional signal. *Phys. Earth Planet. Inter.* **124**, 45–70 (2001).
23. Tanioka, Y. & Satake, K. Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys. Res. Lett.* **23**, 1549–1552 (1996).
24. Thatcher, W. Earthquake recurrence and risk assessment in circum-Pacific seismic gaps. *Nature* **341**, 432–434 (1989).
25. Schwartz, S. Y. Noncharacteristic behavior and complex recurrence of large subduction zone earthquakes. *J. Geophys. Res.* **104**, 23111–23125 (1999).
26. Ando, M. Source mechanisms and tectonic significance of historical earthquakes along the Nankai trough. *Japan. Tectonophysics* **27**, 119–140 (1975).
27. Mansinha, L. & Smylie, D. E. The displacement fields of inclined faults. *Bull. Seismol. Soc. Am.* **61**, 1433–1440 (1971).
28. Imamura, F. in *Long-wave Runup Models* (eds Yeh, H., Liu, P. & Synolakis, C.) 25–42 (World Scientific, Singapore, 1996).
29. Satake, K. Linear and nonlinear computations of the 1992 Nicaragua earthquake tsunami. *Pure Appl. Geophys.* **144**, 455–470 (1995).
30. Iwasaki, T. & Mano, A. Two-dimensional numerical simulation of tsunami run-ups in the Eulerian description [in Japanese]. *Proc. 26th Conf. Coastal Eng. JSCE*, 70–72 (Japan Soc. Civil Engineers, Tokyo, 1979).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Hirakawa, Y. Nishimura, Y. Sawai, A. Moore, H. Kelsey and E. Hemphill-Haley for guidance in the field, and F. Akiba for diatom analysis. We also thank R. Stein, J. Bourgeois, K. Shimazaki, S. Toda and H. Kelsey for comments on the manuscript. The work formed part of the US–Japan Common Agenda for earthquake and tsunami studies.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K. Satake (kenji.satake@aist.go.jp).

The hydrodynamics of water strider locomotion

David L. Hu¹, Brian Chan² & John W. M. Bush¹

¹Department of Mathematics, and

²Department of Mechanical Engineering MIT, Cambridge, Massachusetts 02139, USA

Water striders *Gerridae* are insects of characteristic length 1 cm and weight 10 dynes that reside on the surface of ponds, rivers, and the open ocean^{1–4}. Their weight is supported by the surface tension force generated by curvature of the free surface^{5,6}, and they propel themselves by driving their central pair of hydrophobic legs in a sculling motion^{7,8}. Previous investigators have assumed that the hydrodynamic propulsion of the water strider relies on momentum transfer by surface waves^{1,9,10}. This assumption leads to Denny's paradox¹¹: infant water striders, whose legs are too slow to generate waves, should be incapable of propelling themselves along the surface. We here resolve this paradox through reporting the results of high-speed video and particle-tracking studies. Experiments reveal that the strider transfers momentum to the underlying fluid not primarily through capillary waves, but rather through hemispherical vortices shed by its driving legs. This insight guided us in constructing a self-contained mechanical water strider whose means of propulsion is analogous to that of its natural counterpart.

Whereas substantial advances have been made in the field of biolocomotion^{9,12–16}, the hydrodynamics underlying the surface locomotion of semiaquatic insects remains poorly understood^{1,9,11,17}. There are two means of walking on water according to the relative magnitudes of the body weight Mg and the maximum

curvature force σP , where M is the body mass, g the gravitational acceleration, σ the surface tension and P the contact perimeter of the water-walker¹⁷. Water-walkers characterized by $M_c = \frac{Mg}{\sigma P} > 1$, such as the basilisk lizard, rely on the force generated by their feet slapping the surface and propelling water downward¹⁸. Creatures with $M_c < 1$, such as the water strider may reside at rest on the surface, supported by the curvature force generated by distortion of the free surface (Fig. 1). The body and legs of the water strider are covered by thousands of hairs¹ that render its legs effectively non-wetting¹⁹. During their rowing stroke, water striders drive their middle legs against the water without penetrating the surface and may achieve peak speeds of 150 cm s^{-1} . The striders may launch themselves with a vertical component, or glide along the surface.

The force balance on a stationary water strider may be understood in terms of the statics of floating, non-wetting bodies⁶. The weight of a stationary water strider is supported by some combination of the buoyancy force, F_b , and the curvature force, F_c , associated with the surface tension σ : $Mg = F_b + F_c$. F_b is deduced by integrating the hydrostatic pressure over the body area in contact with the water, while F_c is deduced by integrating the curvature pressure over this area, or equivalently the vertical component of the surface tension, $\sigma \sin \theta$, along the contact perimeter (Fig. 1b). Keller⁶ demonstrated that the ratio of F_b to F_c is equal to that of the fluid volumes displaced inside and outside the contact line. For a long thin water-strider leg, this ratio is $F_b/F_c \sim w/L_c \ll 1$, where $w \approx 40 \mu\text{m}$ is the leg radius¹, $L_c = (\sigma/\rho g)^{1/2} \approx 2 \text{ mm}$ the capillary length, and ρ the density of water. The strider's weight is supported almost exclusively by surface tension. Figure 2 illustrates the dependence of the maximum surface tension force on body weight for 342 species of water striders. The observed dependence illustrates that as the striders increase in size, their legs become proportionately longer. It is only thus that the largest water strider (marked C in Fig. 2), *Gigantometra gigas*²⁰, whose length may exceed 20 cm, is a viable water-walker.

For the water strider to move, Newton's third law requires that it transfer momentum to the underlying fluid. It has previously been assumed that capillary waves are the sole means by which to accomplish this momentum transfer^{1,9–11}. Denny⁹ suggested that the leg speed of the infant water strider is less than the minimum phase speed of surface waves²¹, $c_m = (4g\sigma/\rho)^{1/4} \approx 23.2 \text{ cm s}^{-1}$; consequently, the infants are incapable of generating waves and so transferring momentum to the underlying fluid. According to this physical picture, infant water striders cannot swim, an inference referred to as Denny's paradox^{9,11}.

A series of laboratory experiments were conducted in order to elucidate the hydrodynamic propulsion mechanism of the water strider *Gerris remigis*. Water striders were collected from local freshwater ponds and maintained in aquaria. The striders reproduced every several weeks, providing the opportunity to study the first-instar nymph water striders in a laboratory setting. The striders were filmed using a high-speed video camera at 500 frames per second and the images were digitized and analysed using Midas motion analysis software. Particle tracking was performed by seeding water with either Kalliroscope or Pliolite particles of size 50–100 μm . Dye studies were performed using food colouring and thymol blue. Images of the waves and vortices generated by the strider motion were obtained from both plan and side views; their form and speed were measured directly following the stroke. The surface signature of the capillary waves was measured with a technique adopted from Schooley²².

The Reynolds number characterizing the adult leg stroke is $\text{Re} = UL_2/\nu \approx 10^3$, where $U \approx 100 \text{ cm s}^{-1}$ is the peak leg speed and $L_2 \approx 0.3 \text{ cm}$ is the length of the rowing leg's tarsal segment (see Figs 1b and 2), which prescribes the size of the dynamic meniscus forced by the leg stroke²³. For the 0.01-s duration of the stroke, the driving legs apply a total force $F \approx 50$ dynes, the magnitude of

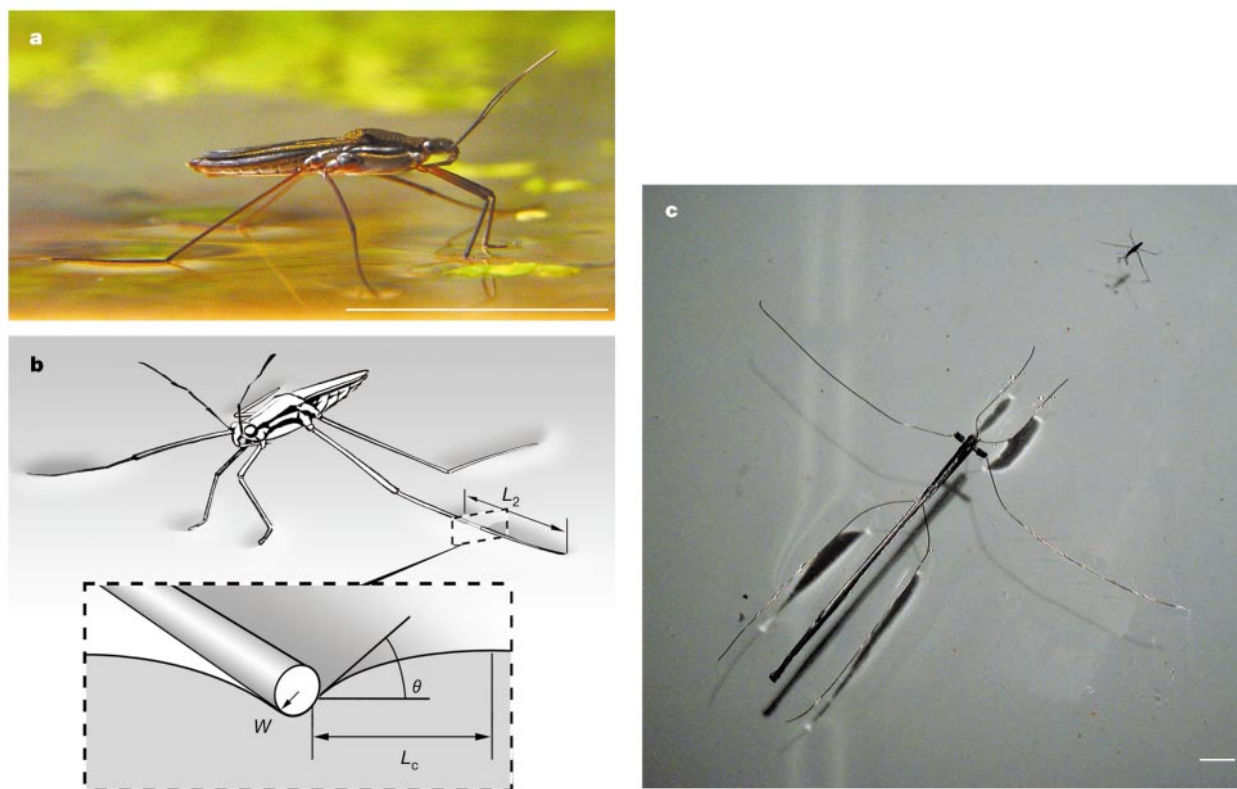


Figure 1 Natural and mechanical water striders. **a**, An adult water strider *Gerris remigis*. **b**, The static strider on the free surface, distortion of which generates the curvature force per unit leg length $2\sigma \sin \theta$ that supports the strider's weight. **c**, An adult water strider facing its mechanical counterpart. Robostrider is 9 cm long, weighs 0.35 g, and has proportions consistent with those of its natural counterpart. Its legs, composed of 0.2-mm

gauge stainless steel wire, are hydrophobic and its body was fashioned from lightweight aluminium. Robostrider is powered by an elastic thread (spring constant $310 \text{ dynes cm}^{-1}$) running the length of its body and coupled to its driving legs through a pulley. The resulting force per unit length along the driving legs is 55 dynes cm^{-1} . Scale bars, 1 cm.

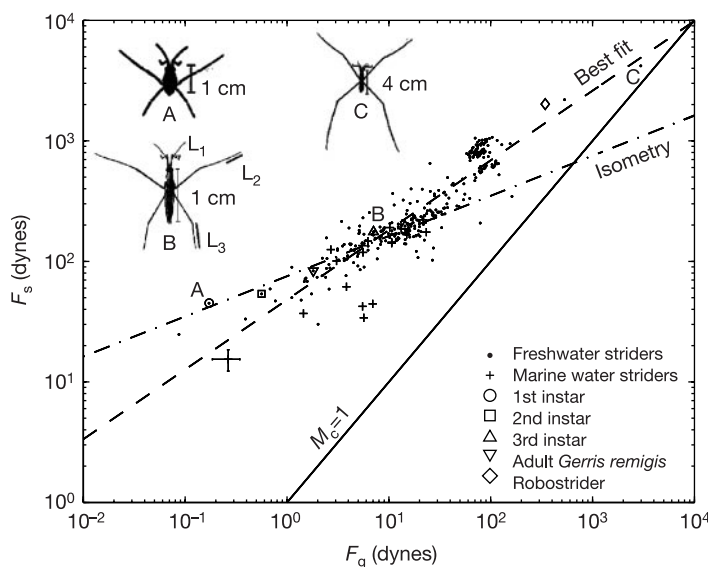


Figure 2 The relation between maximum curvature force $F_s = \sigma P$ and body weight $F_g = Mg$ for 342 species of water striders. σ is the surface tension of either pond water (67 dynes cm^{-1}) or sea water¹⁷ (78 dynes cm^{-1}) at 14°C and $P = 4(L_1 + L_2 + L_3)$ is twice the combined lengths of the tarsal segments (see strider B). Anatomical measurements were compiled from existing data^{20,26–29}. Open symbols denote striders observed in our laboratory. Insets show the adult *Gerris remigis* (B) and extremes in size: the first-instar infant *Gerris remigis* (A) and the *Gigantometra gigas*²⁰ (C). The solid line

represents $M_c = 1$, the minimum requirement for static stability on the surface. The surface tension force is more than adequate to support the water strider's weight; however, the margin of safety (the distance above $M_c = 1$) decreases with increasing body size. If the proportions of the water strider were independent of its characteristic size L , one would expect $P \sim L$ and hence $F_s \sim L$, and $F_g \sim L^3$; isometry would thus suggest $F_s \sim F_g^{1/3}$, a relation indicated by the dash-dotted line. The best fit to the data is given by $F_s = 48F_g^{0.58}$ (dashed line). Characteristic error bars are shown.

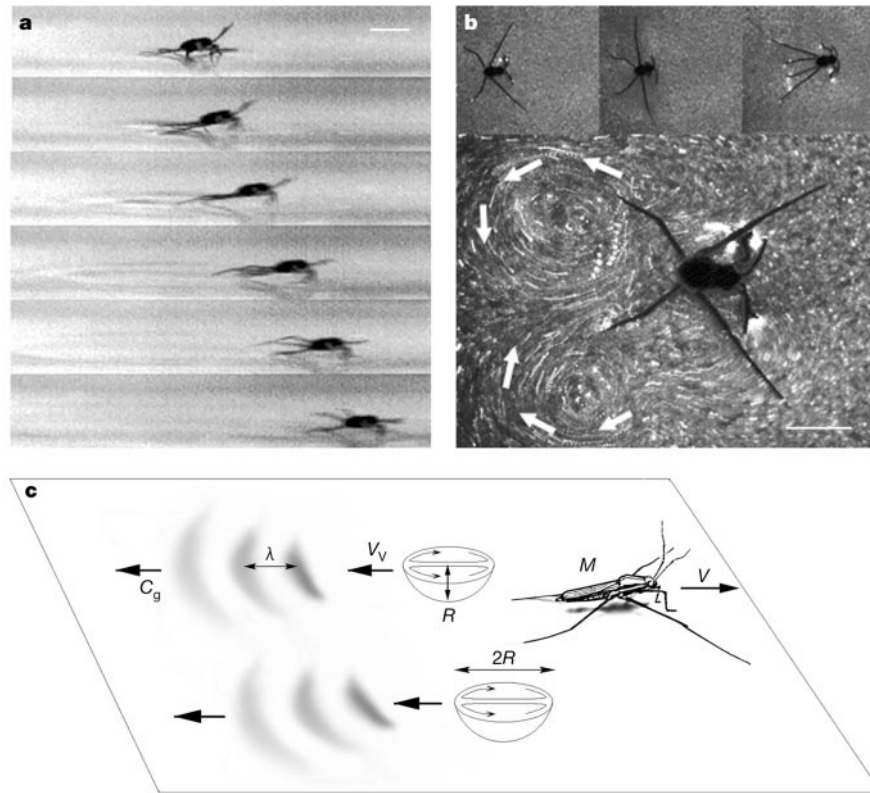


Figure 3 The flow generated by the driving stroke of the water strider. **a, b.** The stroke of a one-day-old first-instar water strider. Sequential images were taken 0.016 s apart. **a,** Side view. Note the weak capillary waves evident in its wake. **b,** Plan view. The underlying flow is rendered visible by suspended particles. For the lowermost image,

fifteen photographs taken 0.002 s apart were superimposed. Note the vortical motion in the wake; the flow direction is indicated. The strider legs are cocked for the next stroke. Scale bars, 1 mm. **c.** A schematic illustration of the flow structures generated by the driving stroke: capillary waves and subsurface hemispherical vortices.

which was deduced independently by measuring the strider's acceleration and leaping height. The applied force per unit length along its driving legs is thus approximately $50/0.6 \approx 80 \text{ dynes cm}^{-1}$. An applied force per unit length in excess of $2\sigma \approx 140 \text{ dynes cm}^{-1}$ will result in the strider penetrating the free surface. The water strider is thus ideally tuned to life at the water surface: it applies as great a force as possible without jeopardizing its status as a water-walker.

The propulsion of a one-day-old first-instar is detailed in Fig. 3. Particle tracking reveals that the infant strider transfers momentum to the fluid through dipolar vortices shed by its rowing motion. The wake of the adult water strider is similarly marked by distinct vortex dipole pairs that translate backwards at a characteristic speed $V_v \approx 4 \text{ cm s}^{-1}$ (Figs 3c and 4). Video images captured from a side view indicate that the dipolar vortices are roughly hemispherical, with a characteristic radius $R \approx 0.4 \text{ cm}$. The vertical extent of the hemispherical vortices greatly exceeds the static meniscus depth²⁴, $120 \mu\text{m}$, but is comparable to the maximum penetration depth of the meniscus adjoining the driving leg, 0.1 cm . A strider of mass $M \approx 0.01 \text{ g}$ achieves a characteristic speed $V \approx 100 \text{ cm s}^{-1}$ and so has a momentum $P = MV \approx 1 \text{ g cm s}^{-1}$. The total momentum in the pair of dipolar vortices of mass $M_v = 2\pi R^3/3$ is $P_v = 2M_v V_v \approx 1 \text{ g cm s}^{-1}$, and so comparable to that of the strider.

The leg stroke may also produce a capillary wave packet, whose contribution to the momentum transfer may be calculated. We consider linear monochromatic deep-water capillary waves with surface deflection $\zeta(x,t) = ae^{i(kx-\omega t)}$ propagating in the x -direction with a group speed $c_g = d\omega/dk$, phase speed $c = \omega/k$, amplitude a , wavelength $\lambda = 2\pi/k$ and lateral extent W . The time-averaged horizontal momentum associated with a single wavelength,

$P_w = \pi\sigma ka^2 Wc^{-1}$, may be computed from the velocity field and relations between wave kinetic energy and momentum^{21,25}. Our measurements indicate that the leg stroke typically generates a wave train consisting of three waves with characteristic wavelength $\lambda \approx 1 \text{ cm}$, phase speed $c \approx 30 \text{ cm s}^{-1}$, amplitude $a \approx 0.01$ – 0.05 cm , and width $L_2 \approx 0.3 \text{ cm}$ (see Fig. 3c). The net momentum carried by the capillary wave packet thus has a maximum value $P_w \approx 0.05 \text{ g cm s}^{-1}$, an order of magnitude less than the momentum of the strider.

The momentum transported by vortices in the wake of the water strider is comparable to that of the strider, and greatly in excess of that transported in the capillary wave field; moreover, the striders are capable of propelling themselves without generating discernible capillary waves. We thus conclude that capillary waves do not play an essential role in the propulsion of *Gerridae*, and thereby circumvent Denny's paradox. The strider generates its thrust by rowing, using its legs as oars and its menisci as blades. As in the case of rowing boats, while waves are an inevitable consequence of the rowing action, they do not play a significant role in the momentum transfer necessary for propulsion. We note that their mode of propulsion relies on the Reynolds number exceeding a critical value of approximately 100, suggesting a bound on the minimum size of water striders. Our continuing studies of water strider dynamics will follow those of birds, insects and fish^{11,15,16} in characterizing the hydrodynamic forces acting on the body through detailed examination of the flows generated during the propulsive stroke.

We designed a mechanical water strider, Robostrider, constructed to mimic the motion of a water strider (Fig. 1c). Its proportions and M_c value were consistent with those of its natural counterpart

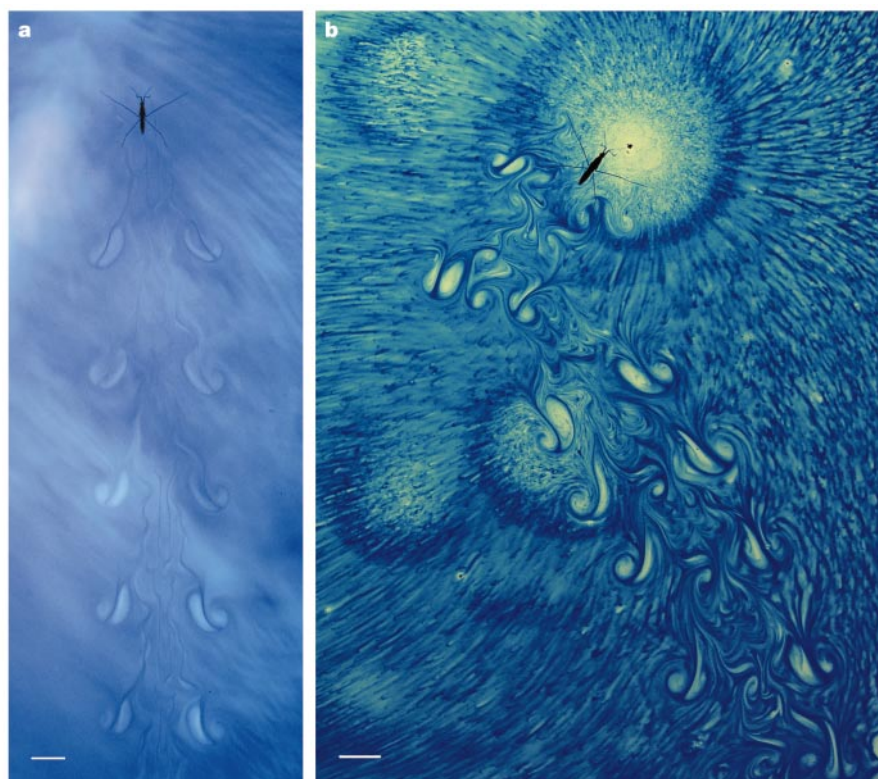


Figure 4 Dipolar vortices in the wake of the adult water strider. Images captured from a side view indicate their hemispherical form. **a**, A thin layer (2–5 mm) of thymol blue was established on the surface of the water, disturbance of which revealed the vortical footprints of the water strider. **b**, The ambient texture results from Marangoni convection³⁰ in the suspending fluid prompted by thymol blue on its surface. The starburst pattern

results from the chunk of thymol blue evident at its centre reducing the local surface tension, thus driving surface divergence that sweeps away the dyed surface layer. The fluid is illuminated from below; consequently, the light-seeking water strider is drawn to the starbursts. Scale bars, 1 cm.

(Fig. 2). The challenge was constructing a self-contained device sufficiently light to be supported by surface tension and capable of rowing without breaking the water surface. An important design criterion, that the force per unit length along the driving legs not exceed 2σ , was met by appropriate choice of elastic thread and pulley. High-speed video footage indicates that Robostrider does not break the surface despite leg speeds of 18 cm s^{-1} . Like its natural counterpart, the Robostrider generates both capillary waves and vortices, and the principal momentum transfer is in the form of vortices shed by the rowing motion. Robostrider travels half a body length per stroke in a style less elegant than its natural counterpart. □

Received 8 February; accepted 6 May 2003; doi:10.1038/nature01793.

- Andersen, N. M. A comparative study of locomotion on the water surface in semiaquatic bugs (Insecta, Hemiptera, Gerromorpha). *Vidensk. Meddr. Dansk. Naturh. Foren.* **139**, 337–396 (1976).
- Brinkhurst, R. O. Studies on the functional morphology of *Gerris najas degeer* (Hem. Het. Gerridae). *Proc. Zool. Soc. Lond.* **133**, 531–559 (1960).
- Murphey, R. K. A. Motor control of orientation to prey by the water strider *Gerris remigis*. *Z. Vergl. Physiol.* **72**, 150–167 (1971).
- Wilcox, R. S. Sex discrimination in *Gerris remigis*: Role of a surface wave signal. *Science* **206**, 1325–1327 (1979).
- Baudoin, R. La physico-chimie des surfaces dans la vie des arthropodes aériens des miroirs d'eau, des rivières marines et lacustres et de la zone intercotidale. *Bull. Biol. Fr. Belg.* **89**, 16–164 (1955).
- Keller, J. B. Surface tension force on a partly submerged body. *Phys. Fluids* **10**, 3009–3010 (1998).
- Darnhofer-Demar, B. Zur Fortbewegung des Wasserläufers *Gerris lacustris* L. auf der Wasseroberfläche. *Zool. Anz. Suppl.* **32**, 430–439 (1969).
- Bowdan, E. Walking and rowing in the water strider, *Gerris remigis*. *J. Comp. Physiol.* **123**, 43–49 (1978).
- Denny, M. W. *Air and Water: The Biology and Physics of Life's Media* (Princeton Univ. Press, Princeton, NJ, 1993).
- Sun, S. M. & Keller, J. B. Capillary-gravity wave drag. *Phys. Fluids* **13**, 2146–2151 (2001).
- Suter, R. B., Rosenberg, O., Loeb, S., Wildman, H. & Long, J. H. Locomotion on the water surface: propulsive mechanisms of the fisher spider *Dolomedes triton*. *J. Exp. Biol.* **200**, 2523–2538 (1997).
- Childress, S. *Mechanics of Swimming and Flying* (Cambridge Univ. Press, Cambridge, UK, 1981).
- Vogel, S. *Life's Devices* (Princeton Univ. Press, Princeton, NJ, 1988).

- Dickinson, M. H. et al. How animals move: an integrated view. *Science* **288**, 100–106 (2000).
- Rayner, J. M. V., Jones, G. & Thomas, A. Vortex flow visualizations reveal change in upstroke function with flight speed in bats. *Nature* **321**, 162–164 (1986).
- Ellington, C. P. Oxygen consumption of bumblebees in forward flight. *Nature* **347**, 472–473 (1990).
- Vogel, S. *Life in Moving Fluids* (Princeton Univ. Press, Princeton, NJ, 1994).
- Glasheen, J. W. & McMahon, T. A. A hydrodynamic model of locomotion in the Basilisk Lizard. *Nature* **380**, 340–342 (1996).
- de Gennes, P.-G., Brochard-Wyart, F. & Quere, D. *Gouttes, Bulles, Perles et Ondes* (Belin, Collection Echelles, Paris, 2002).
- Tseng, M. & Rowe, L. Sexual dimorphism and allometry in the giant water strider *Gigantometra gigas*. *Can. J. Zool.* **77**, 923–929 (1999).
- Lamb, H. *Hydrodynamics*, 6th edn (Cambridge Univ. Press, Cambridge, 1932).
- Schooley, A. H. Profiles of wind-created water waves in the capillary-gravity transition region. *J. Mar. Res.* **16**, 100–108 (1958).
- Suter, R. B. & Wildman, H. Locomotion on the water surface: hydrodynamic constraints on rowing velocity require a gait change. *J. Exp. Biol.* **202**, 2771–2785 (1999).
- Matsuda, K., Watanabe, S. & Eiju, T. Real-time measurement of large liquid surface deformation using a holographic shearing interferometer. *Appl. Opt.* **24**, 4443–4447 (1985).
- Starr, V. P. Momentum and energy integrals for gravity waves of finite height. *J. Mar. Res.* **6**, 175–193 (1947).
- Andersen, N. M. *The Semiaquatic Bugs (Hemiptera, Gerromorpha): Phylogeny, Adaptations, Biogeography and Classification* (Scandinavian Science, Klampenborg, 1982).
- Hungerford, H. B. & Matsuda, R. Keys to subfamilies, tribes, genera and subgenera of the Gerridae of the world. *Kans. Univ. Sci. Bull.* **41** (1960).
- Henry, T. J. & Froeschner, R. C. (eds) *Catalog of the Heteroptera, of True Bugs, of Canada and the Continental United States* (E. J. Brill, Leiden, 1998).
- Cobben, R. H. The Hemiptera of the Netherlands. *Stud. Fauna Curacao Caribb. Islands* **11**, 1–34 (1960).
- Scriven, L. E. & Sterling, C. V. The Marangoni effects. *Nature* **187**, 186–188 (1970).

Acknowledgements We thank A. Chau for preparing Fig. 2, M. Hancock, M. Shelley and R. Rosales for discussions, and MIT's Edgerton Center for lending us their high-speed video equipment. J.W.M.B. gratefully acknowledges the financial support of the NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.B. (bush@math.mit.edu).

Bottlenose dolphins perceive object features through echolocation

Heidi E. Harley^{1,2}, Erika A. Putman² & Herbert L. Roitblat³

¹New College of Florida, Division of Social Sciences, 5700 N. Tamiami Trail, Sarasota, Florida 34243, USA

²Animal Programs, Walt Disney World Resort, Epcot's The Living Seas, EC TRL W-251, P.O. Box 10,000, Lake Buena Vista, Florida 32830, USA

³DolphinSearch, Inc., 474 E. Main Street, Ventura, California 93001, USA

How organisms (including people) recognize distant objects is a fundamental question¹. The correspondence between object characteristics (distal stimuli), like visual shape, and sensory characteristics (proximal stimuli), like retinal projection, is ambiguous. The view that sensory systems are 'designed' to 'pick up' ecologically useful information is vague about how such mechanisms might work². In echolocating dolphins, which are studied as models for object recognition sonar systems, the correspondence between echo characteristics and object characteristics is less clear³. Many cognitive scientists assume that object characteristics are extracted from proximal stimuli, but evidence for this remains ambiguous. For example, a dolphin may store 'sound templates' in its brain and identify whole objects by listening for a particular sound. Alternatively, a dolphin's brain may contain algorithms, derived through natural endowments or experience or both, which allow it to identify object characteristics based on sounds. The standard method used to address this question in many species^{4–7} is indirect and has led to equivocal results with dolphins^{8–10}. Here we outline an appropriate method and test it to show that dolphins extract object characteristics directly from echoes.

Most data about cetacean echolocation are based on studies in which dolphins used echolocation alone to perform discrimination tasks^{11–15}. The dolphin is trained, for example, to make a particular response in the presence of one stimulus and to make a different response in the presence of others¹⁶. In this case the dolphin may merely be discriminating among sounds (proximal stimuli) rather than extracting object characteristics (distal stimuli). When evidence for the same object characteristic (for example, convexity or texture) is available through more than one sensory system, we can investigate whether organisms detect the object characteristics directly. If they detect object characteristics, then they should be able to detect immediately a correspondence between the evidence for a characteristic perceived by one sensory system (modality) and that perceived by

another. On the other hand, if they are responding solely to the proximal stimuli, then they should require some (nearly) simultaneous experience with the two modalities before they can detect a correspondence. Put concretely, animals that respond solely on the basis of proximal stimuli should be equally likely to learn arbitrary cross-modal pairings, whereas animals that respond to object characteristics directly should be biased to respond to the inherent correspondence of the cross-modal evidence and should find it difficult to ignore this correspondence.

The standard method for discriminating between these two mechanisms is a cross-modal identity matching task: a subject uses one modality (for example, touch) to experience a sample object, and then must use a second modality (for example, vision) to identify an identical object from several alternatives. If the object is unfamiliar, the subject will never have had an opportunity to associate its two different proximal experiences, and thus correct matches imply a recognition that the same object characteristic was identified by each of the sensory systems. However, only first trials with unfamiliar objects can be used in this type of analysis, because thereafter the subject has had the opportunity to form an association.

Unfortunately, reliance on first trials can make results from cross-modal matching difficult to interpret, because unfamiliar stimuli can lead to matching failures for many reasons (such as neophobia). For example, it is not clear why one dolphin, well versed in cross-modal matching procedures, could only match objects from two of four completely novel object pairs even though post-test intra-modal matching was strong with all stimuli¹⁰. Analyses of first-trial performance accuracy with this method may also be inflated: common practice allows the inclusion of first-sample presentations as first-trial data even if the sample object has been used as a choice alternative in a previous trial. Thus, if a subject has mastered a win-stay/lose-shift strategy, and/or can learn through exclusion, then a single trial in a two-alternative task could lead to enough information for the subject to perform perfectly thereafter based on association alone.

A more direct test of object characteristic extraction exploits the fact that if cross-modal correspondence is always learned, then it should not matter whether or not the animal is rewarded for choosing the alternative that actually corresponds to the sample or one that is arbitrarily selected by the experimenter to be the correct choice. If all different sensory experiences of the same object are arbitrarily linked, then the fact that the same object is presented to both sensory systems is not perceptible. Arbitrary relations would then be as easy to learn as actual feature correspondence. On the other hand, if the animal is aware of object features, then a recognition of equivalence between different sensory experiences of the same object features should occur. We exploited this dichotomy by modifying the cross-modal matching-to-sample task. Following certain samples, the dolphin was rewarded for choosing the alternative comparison stimulus that did actually match the sample. Following the remaining samples, the dolphin was rewarded for choosing an arbitrary, but consistent, alternative that was different from the sample. Object perception predicts that the dolphin will find the task easier when the designated correct cross-modal comparison item matches the sample than when it



Figure 1 Examples of stimulus objects. The ruler in the lower left corner represents 10 cm.

Table 1 Choice frequencies across sets in which only identity matches were reinforced

	Objects A	Objects B	Objects C
Objects A	440	118	125
Objects B	127	426	130
Objects C	77	111	496

Identity-based choices are presented on the diagonal. Frequencies in bold font indicate reinforced choices. Data are collapsed across 12 3-object sets.

Table 2 Choice frequencies across sets involving arbitrarily selected relations

	Objects A	Objects B	Objects C
Objects A	268	16	52
Objects B	74	230	32
Objects C	67	22	247

Identity-based choices are presented on the diagonal. The first two rows include stimuli that were arbitrarily paired. Frequencies in bold font indicate reinforced choices. Data are collapsed across six 3-object sets.

does not actually match. Proximal perception predicts that the dolphin will find all tasks equally difficult whether or not the designated correct cross-modal comparison object actually matches.

In this study, an adult male bottlenose dolphin at Disney's Epcot's The Living Seas that had extensive previous experience with echoic identity matching¹⁵ was trained to match cross-modally. He was later tested with 54 unfamiliar objects organized into 18 3-object sets. Objects were chosen because they varied on as many dimensions as possible. See Fig. 1 for some examples. Each object set was presented in one of two conditions. In the echoic-visual condition, samples were presented underwater behind a thin black polyethylene screen that was visually opaque but echoically transparent; the choice alternatives were presented in air, thus making them visually, but not echoically, available. (Dolphins cannot echolocate in air, because the impedance mismatch between water and air is too great³.) Conversely, in the visual-echoic condition, samples were presented in air and alternatives were presented underwater behind a black screen. The dolphin was reinforced for arbitrarily paired object choices (that is, for choosing object B after experiencing sample A and vice versa) with 2 objects within 6 of the object sets and for identity matching (for example, for choosing object C after experiencing sample C) for the 42 remaining objects.

Table 1 shows the distribution of responses on those sets in which the animal was rewarded for choosing the matching stimulus. Table 2 shows the corresponding distribution for the tasks in which the dolphin was rewarded for choosing arbitrary stimuli. We used a χ^2 analysis to test the hypothesis that the choice pattern in all tasks was controlled by the reward. If this prediction is true, then the proportion of trials on which the dolphin chose the rewarded alternative should be the same in Table 2 as it is in Table 1. Non-reinforced alternatives are predicted to be chosen indifferently. In contrast to these predictions, the animal tended to choose the matching alternative whether it was reinforced or not (χ^2 ($n = 1008$, d.f. = 3) = 1654.83, $P < 0.0001$). In the cross-modal conditions with the arbitrarily paired objects, the dolphin made 498/672 (74.1%) identity-based choices versus 90/672 (13.4%) rewarded choices. The dolphin chose the third alternative in the subset the remainder of the time. Three choices were always available; therefore chance performance was 33%.

Even when the dolphin was rewarded for choosing an arbitrary stimulus, it still chose the identical object. Therefore, reinforcement cannot be the mechanism that allows the dolphin to recognize the correspondence between the different sensory modalities. Further, because the stimuli were never presented contemporaneously to the two modalities, the dolphin never had the experience with the specific stimuli that would allow it to form an association. Instead, to recognize that the two sensory modalities both represented the same object, the dolphin had to have the knowledge of the object characteristics that they signalled. Organisms can perceive object characteristics directly. □

Methods

More information about the environment at Disney's The Living Seas (a 5.7-million-gallon circular tank housing many aquatic species) and the experimentally experienced adult male dolphin (*Tursiops truncatus*) subject, Toby, is available elsewhere¹⁵.

Materials

The stimuli were 54 assorted 'junk' or hardware-store objects that were unfamiliar to the dolphin. The largest object was 34.29 cm × 34.29 cm; the smallest was 5.08 cm × 7.62 cm. Objects were made from rubber, hard plastic, air-filled plastic, Styrofoam, metal and wood. The objects were organized into 3-item matching-to-sample problems. Three 3-item problems, consistently grouped, were presented in each 54-trial session. (Figures of all objects are shown in Supplementary Information.)

Echoic samples were presented in the echoic sample apparatus, a 0.7 m × 0.7 m PVC-framed box tightly covered in 0.1 mm black polyethylene that was echoically transparent but visually opaque. Echoic comparison stimuli were presented in the echoic choice apparatus, a larger (front: 2.76 m × 0.90 m; sides: 0.98 m × 0.90 m) but similarly fashioned rectangular structure used in previous studies¹⁵, the top of which was positioned 1 m below the water's surface. Visual samples were presented in air in front of a black plastic screen (0.59 m × 0.46 m). Visual comparison stimuli were presented in air, behind acrylic windows in an underwater room accessible to the public and visible from the tank.

Procedure

Stimuli in this experiment were presented in two cross-modal conditions. In the visual-echoic (V-E) condition, a sample was presented in air 0.25 m above the water's surface, thus making the sample visually, but not echoically, accessible; the three choice alternatives were presented behind the echoic choice apparatus, making the choices echoically, but not visually, accessible. (To confirm that the dolphin could not echolocate the visual samples in air, we ran a single control session with three objects in which samples were presented in air either in front of or behind the screen. The dolphin responded appropriately in the first instance and refused to make choices in the second, thus suggesting that he had not been exposed to the sample when it was in air and behind the screen.) In the E-V condition, the sample was presented underwater in the echoic sample apparatus, making the sample echoically, but not visually, accessible; the choice alternatives were hung in air behind an acrylic window, thus making them visually, but not echoically, accessible.

At the beginning of each V-E trial, the dolphin was positioned vertically in the water so as to view the sample as long as he wished. After viewing he swam to the echoic choice array located several metres behind him. He then positioned himself in front of one object. After about 3 s, an assistant who was naive to the sample's identity reported the dolphin's choice to the trainer who blew a whistle and reinforced him with two small fish for a correct choice, or tapped on a metal platform to recall him for an incorrect choice. E-V trials were identical except for the method of stimulus presentation. Inter-trial intervals averaged 62 s.

Within a session, each stimulus was presented as the sample object an equal number of times. Each object was also presented equally often in each choice position. The order of the trials was randomized. The dolphin experienced the first three object sets in the E-V condition, the second three sets in the V-E condition, the third in the E-V condition, the fourth in the V-E condition, and so on. The number of 54-trial sessions varied from 8–10 with each set, because in three cases objects were accidentally dropped into the pool before the final sessions, thereby precluding further use in the study because the dolphin may have had the opportunity to both see and echolocate them simultaneously. The objects were never available simultaneously to both modalities during the study. Reinforcement occurred for identity- and arbitrarily matched pairs as previously described.

Received 26 November 2002; accepted 9 June 2003; doi:10.1038/nature01846.

- Goldstein, E. B. (ed.) *Blackwell Handbook of Perception* (Blackwell, Oxford, 2001).
- Gibson, J. J. *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston, 1979).
- Au, W. W. L. *The Sonar of Dolphins* (Springer, New York, 1993).
- Garbin, C. P. Visual-haptic perceptual nonequivalence of shape information and its impact upon cross-modal performance. *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 547–553 (1988).
- Davenport, R. K. in *Origins and Evolution of Language and Speech* (eds Harnad, S. R., Steklis, H. D. & Lancaster, J.). *Ann. NY Acad. Sci.* **280**, 143–149 (1976).
- Gunderson, V. M., Rose, S. A. & Grant-Webster, K. S. Cross-modal transfer in high- and low-risk infant pigtailed macaque monkeys. *Dev. Psychol.* **26**, 576–581 (1990).
- Weiskrantz, L. & Cowey, A. Cross-modal matching in the rhesus monkey using a single pair of stimuli. *Neuropsychologia* **13**, 257–261 (1975).
- Harley, H. E., Roitblat, H. L. & Nachtigall, P. E. Object representation in the bottlenose dolphin (*Tursiops truncatus*): Integration of visual and echoic information. *J. Exp. Psychol. Anim. Behav. Process.* **22**, 164–174 (1996).
- Pack, A. A. & Herman, L. M. Sensory integration in the bottlenose dolphin: Immediate recognition of complex shapes across the senses of echolocation and vision. *J. Acoust. Soc. Am.* **98**, 722–733 (1995).
- Herman, L. M., Pack, A. A. & Hoffman-Kuhnt, M. Seeing through sound: Dolphins (*Tursiops truncatus*) perceive the spatial structure of objects through echolocation. *J. Comp. Psychol.* **112**, 292–305 (1998).
- Evans, W. E. Echolocation by marine dolphins and one species of fresh-water dolphin. *J. Acoust. Soc. Am.* **54**, 191–199 (1973).
- Au, W. W. L., Schusterman, R. J. & Kersting, D. A. in *Animal Sonar Systems* (eds Busnel, R. G. & Fish, J. F.) 859–862 (Plenum Press, New York, 1980).
- Nachtigall, P. E., Murchison, A. E. & Au, W. W. L. in *Animal Sonar Systems* (eds Busnel, R. G. & Fish, J. F.) 945–947 (Plenum Press, New York, 1980).
- Roitblat, H. L., Penner, R. H. & Nachtigall, P. E. Matching-to-sample by an echolocating dolphin. *J. Exp. Psychol. Anim. Behav. Process.* **16**, 85–95 (1990).
- Xitco, M. J. Jr & Roitblat, H. L. Object recognition through eavesdropping: Passive echolocation in bottlenose dolphins. *Anim. Learn. Behav.* **24**, 355–365 (1996).
- Au, W. W. L. & Pawloski, D. A. Cylinder wall thickness discrimination by an echolocating dolphin. *J. Comp. Physiol. A* **172**, 41–47 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by Walt Disney Company's Animal Programs, New College of Florida and the University of South Florida. All experimental procedures were evaluated and approved according to animal welfare regulations specified under National Institutes of Health guidelines. We thank: G. Bauer, J. Gory, A. Stamper, W. Fellner and J. Orth for discussion and manuscript review; D. Clark, D. Bickel, W. Fellner and K. Yannessia for research session and data management; J. Davis, A. Stamper, T. Hopkins, B. Stevens, J. Ogden and J. Mellen for administrative support; and M. Muraco, M. Barringer, B. Cavanaugh, L. Davis, D. Feuerbach, C. Goonen, L. Larsen-Plott, K. Odell and C. Litz for training the dolphin.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.E.H. (harley@ncf.edu).

Neuronal populations and single cells representing learned auditory objects

Timothy Q. Gentner & Daniel Margoliash

Department of Organismal Biology and Anatomy, University of Chicago,
1027 East 57th Street, Chicago, Illinois 60637, USA

The neural representations associated with learned auditory behaviours, such as recognizing individuals based on their vocalizations, are not well described. Higher vertebrates learn to recognize complex conspecific vocalizations that comprise sequences of easily identified, naturally occurring auditory objects^{1,2}, which should facilitate the analysis of higher auditory pathways. Here we describe the first example of neurons selective for learned conspecific vocalizations in adult animals—in starlings that have been trained operantly to recognize conspecific songs. The neuronal population is found in a non-primary forebrain auditory region, exhibits increased responses to the set of learned songs compared with novel songs, and shows differential responses to categories of learned songs based on recognition training contingencies. Within the population, many cells respond highly selectively to a subset of specific motifs (acoustic objects) present only in the learned songs. Such neuronal selectivity may contribute to song-recognition behaviour, which in starlings is sensitive to motif identity^{3,4}. In this system, both top-down and bottom-up processes may modify the tuning properties of neurons during recognition learning, giving rise to plastic representations of behaviourally meaningful auditory objects.

The constraints on natural-object representation imposed by stimulus complexity and behaviour argue for the use of model systems in which both the form and function of the sensory input are well understood². We are investigating the representation of vocal-communication signals in songbirds, in which the ecological relevance of sensory signals (that is, songs) is well established. The competitive reproductive environments in which songbirds function require adult birds to discriminate and classify conspecific vocalizations on the basis of intra-species acoustic variation, and various forms of vocal recognition are widespread among songbirds⁵. Song-recognition behaviour has been studied extensively in European starlings (*Sturnus vulgaris*). Starling songs comprise sequences of repeated, discrete, multiple-note clusters, referred to as 'motifs'⁶ (see, for example, Fig. 3a). The behavioural data suggest that motifs are perceived as auditory objects. Motifs vary in length from 200 to 1,000 ms, but the pattern of notes within a given motif type is largely stable between renditions. Individual birds tend to sing unique motifs. In laboratory experiments with artificial mixtures of songs, the relative proportions of motifs from different

singers predicted song-recognition behaviour, suggesting that starlings form memories of motifs as part of the song-recognition process^{3,4,7}. Recent results have identified auditory-forebrain pathways in songbirds beyond the primary thalamo-recipient telencephalic region known as field L⁸, including ventral hyperstriatal neurons with complex response properties^{9,10} (Fig. 1a). Here we report on data recorded from single neurons in the caudomedial ventral hyperstriatum (cmHV) of starlings trained operantly to recognize sets of conspecific songs (Fig. 1b–d). Many of these neurons responded selectively to learned motifs.

Eight adult starlings were trained, four each on a two-alternative choice or go/no-go procedure (Fig. 1b), to discriminate accurately between four to ten 10-s samples of conspecific songs divided equally into two training sets. Each bird achieved asymptotic behaviour with highly accurate recognition of the training songs (Fig. 1c, d). We then anaesthetized each subject with urethane and recorded the extracellular responses of single neurons in the cmHV to a large ensemble of auditory stimuli. The stimulus ensemble comprised the song stimuli that were used during operant recognition training ('familiar' songs), novel conspecific song stimuli ('unfamiliar' songs) and two synthetic stimuli. The stimulus ensemble was similar across subjects, except that the familiarity or novelty of the song stimuli varied systematically on the basis of their operant training; the unfamiliar songs for one subject served as half of the training songs for another subject (see Methods). Neurons were tested with 33–77 unique song motifs (73–178 total motifs), depending on the set of familiar songs used to train each animal and the set of unfamiliar songs used during testing. Because preliminary experiments indicated that many cmHV neurons had very low spontaneous rates and responded to only a few stimuli, we used the entire stimulus ensemble (familiar and unfamiliar songs, and synthetic stimuli) to search for auditory units as well as to test the neurons' responses. We report here on the 45 single units in cmHV that showed significant excitatory responses to one or more of the test stimuli (see Methods).

As a population, cmHV neurons responded selectively to the class of familiar songs. The mean response strength (RS; see Methods) was strongly biased towards familiar songs (6.7 ± 1.8) compared with unfamiliar songs (3.9 ± 1.3), and the difference between the normalized RS (z-scores; see Methods) for familiar and unfamiliar songs was highly significant ($F_{(1,43)} = 25.55$, $P < 0.0001$; Fig. 2a). The strong response bias for familiar songs was consistent in animals trained under both the two-alternative choice and go/no-go operant regimes ($F_{(1,43)} = 1.85$, no significant interaction found between stimulus familiarity and operant training; Fig. 2a). Thus, in this paradigm, song-recognition learning shapes the responses of cmHV neurons.

Whereas subjects trained using the two-alternative choice procedure showed no reliable difference between response strengths associated with the two sets of training songs (one-tailed test, $t = -0.67$, not significant; $n = 14$ neurons, 4 birds), those trained with the go/no-go procedure did ($n = 31$ neurons, 4 birds). Songs associated with positive reinforcement (S+ stimuli) elicited significantly stronger responses than those associated with punishment (S− stimuli; Fig. 2b) (two-tailed test, $t = 2.68$, $P < 0.05$). This difference was not the result of an overall failure to respond to the S− stimulus, as the S− stimuli elicited stronger responses than unfamiliar songs (two-tailed test, $t = 2.28$, $P < 0.05$; Fig. 2b). Thus, although animals in both operant regimes learned to discriminate equally well between the sets of training songs (Fig. 1c, d), the reinforcement contingencies specific to each regime (see Methods) had differential effects on the distributions of neuronal responses. These task-specific effects may reflect the use of different strategies to solve each task, and/or the differential cost of incorrect responses under each regime (see Methods and Supplementary Information). In either case, the results suggest that 'top-down' associative processes as well as 'bottom-up' stimulus activation shape the

responses of cmHV neurons, which are linked to song-recognition learning.

The bias for familiar songs indicates that cmHV neurons did not respond equally to all songs in the test ensemble. To examine this in more detail, we quantified each neuron's selectivity as a function of the different test stimuli by expressing the maximum stimulus-elicited RS for a given cell as a selectivity index (SI) score. If the SI score for a cell exceeded the 95th percentile in a simulated distribution of random SI scores (see Methods), the cell was termed 'song-selective' for the song that elicited the maximum RS. Cells for which the SI did not exceed the 95th percentile were classed as 'non-selective'. By this definition, 64.4% (29 out of 45) of the cells in the cmHV gave a selective response to one of the test stimuli. Notably, 27 of the 29 song-selective cells preferred one of the training songs (Fig. 3). This overwhelming bias in the proportion of preferred stimuli towards familiar songs is significantly different from that expected by chance, given the distribution of familiar and unfamiliar stimuli in the test ensemble ($\chi^2 = 12.03$, $P < 0.005$) — that is, after correcting for the bias introduced by searching and testing with greater numbers of familiar than unfamiliar songs (Fig. 2c). Different measures of selectivity resulted in different distributions of selective and non-selective cells, but did not eliminate the strong, significant bias towards familiar songs (see Supplementary Information). Among the non-selective cells, the proportions of pre-

ferred stimuli were also skewed towards familiar songs, but the bias was not significantly different from chance ($\chi^2 = 3.50$, $P = 0.18$). The difference in the distributions of responses to the familiar and unfamiliar songs was not wholly dependent on the response to the preferred song. Even when the preferred song was removed from the analysis, the mean RS associated with the familiar songs remained significantly greater than that for the unfamiliar songs ($F_{(1,43)} = 12.67$, $P < 0.001$). This was observed for non-selective as well as song-selective cells (Fig. 2d), showing that, as a population, cmHV neurons evince selectivity for the class (or a subset) of familiar songs. The strong bias in the selectivity of single neurons for familiar songs resembles biases seen in the song control system for a bird's own song^{11,12}. However, this is the first report of neuronal selectivity for songs of conspecifics that adults have learned to recognize, and obtains for neurons outside the song control system.

Both song-selective and non-selective cells could respond with phasic or tonic patterns of activity, and with high or low spontaneous firing rates (Fig. 4a). However, there was a significant bias for song-selective cells to have more phasic responses (Mann–Whitney U -test, $P < 0.01$) and lower spontaneous firing rates (Mann–Whitney U -test, $P < 0.05$) than non-selective cells. Thus, most cells with near-zero spontaneous firing rates and highly phasic responses were selective. For many of the song-selective cells, responses were restricted to one or a small number of repeated

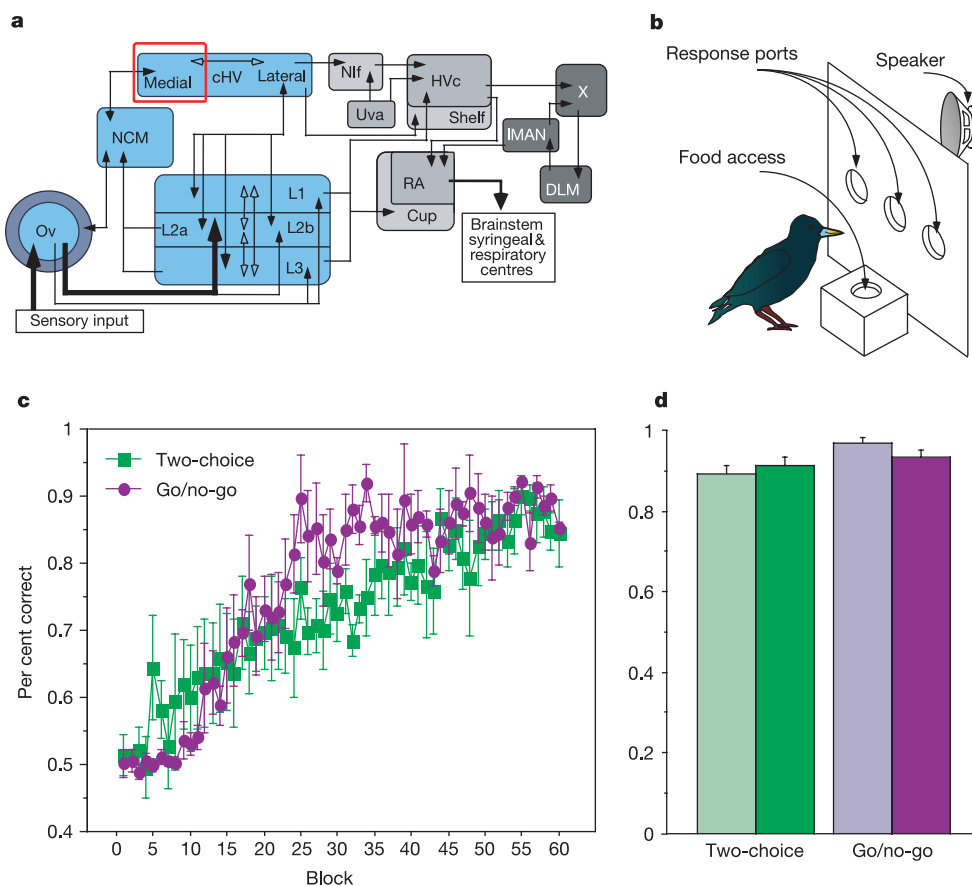


Figure 1 Background and behaviour. **a**, Schematic of the songbird forebrain auditory system (in blue; Ov, nucleus ovoidalis; L1–L3, field L complex; NCM, caudomedial neostriatum; cHV, caudoventral hyperstriatum). The cmHV is outlined in red. Nuclei in the adjacent vocal 'song' system are shown in grey. **b**, Schematic of the operant apparatus. Animals probed openings in the panel in response to different songs to receive a food reward. **c**, Acquisition curves showing mean performance (as the proportion of correct responses) for all subjects over the first 60 blocks of training (100 trials per block).

Acquisition rates differed significantly between training regimes ($F_{(59,354)} = 1.597$, $P < 0.01$). **d**, Mean proportion of correct responses over the last 500 trials before recording, plotted separately for the two sets of training stimuli (light and dark bars) within each regime. The mean ($\pm$ s.e.m.) proportion of correct responses at asymptote (0.93 ± 0.01) was significantly above chance (chi-squared test, $P \ll 0.0001$), and did not vary significantly between the two training regimes, or within regimes between stimulus sets.

motifs within one or a few songs, typically with suppression of background activity for all other motifs (Fig. 3a, b). Examining the response data (RS scores) on a motif-by-motif basis, by simply counting the number of motifs that elicited a significant response from each cell, revealed that the song-selective cells responded to significantly fewer motifs (8.0 ± 1.5) than non-selective cells (20.5 ± 3.8 ; Mann–Whitney U -test, $P < 0.0005$; Fig. 4b). We quantified the relationship between motif features and each neuron's response using a multiple linear regression between the RS associated with each motif in a given cell's test ensemble and the 50 strongest coefficients from a wavelet decomposition of the sonogram for each motif (see Supplementary Information). The mean non-adjusted R^2 value (see Methods) for regressions with the song-selective cells (0.675 ± 0.033) was significantly greater than that for the non-selective cells (0.554 ± 0.048 ; Mann–Whitney U -test, $P < 0.05$), indicating that acoustic features in individual motifs predicted the responses of the song-selective cells better than those of non-selective cells. This neuronal sensitivity to acoustic features is the result of nonlinear response properties, as demonstrated by the poor predictions made by linear estimates of each cell's receptive field (see Supplementary Information). Thus, 'song-selectivity' in cmHV is expressed as a

tendency towards phasic responses to specific acoustic features that appear in only a small number of motifs. The selective representation of acoustic features that are diagnostic for single or small sets of motifs, observed here, could help starlings to evaluate the proportion of familiar motifs within a given song, which behavioural work has shown is important for conspecific song recognition⁴.

The data suggest that mechanisms of experience-dependent representational plasticity act to modify the responses of cmHV neurons on the basis of the functional demands of song recognition. The acoustic features of the motifs (as represented by the wavelet coefficients) were correlated with the response strength of many cmHV neurons. However, the explicit spectrotemporal parameters underlying the selectivity of cmHV neurons have not yet been determined, and it is unlikely that a purely acoustic account of cmHV responses, as implied by the 'feature-detector' terminology, will obtain. The distribution of cmHV response properties depends not only on the spectrotemporal patterns of motifs in familiar and unfamiliar stimuli, but also on the specifics of the conditioning paradigm. All responses to stimuli were rewarded or punished in the two-alternative choice condition, whereas only some were rewarded or punished in the go/no-go condition. Because song recognition

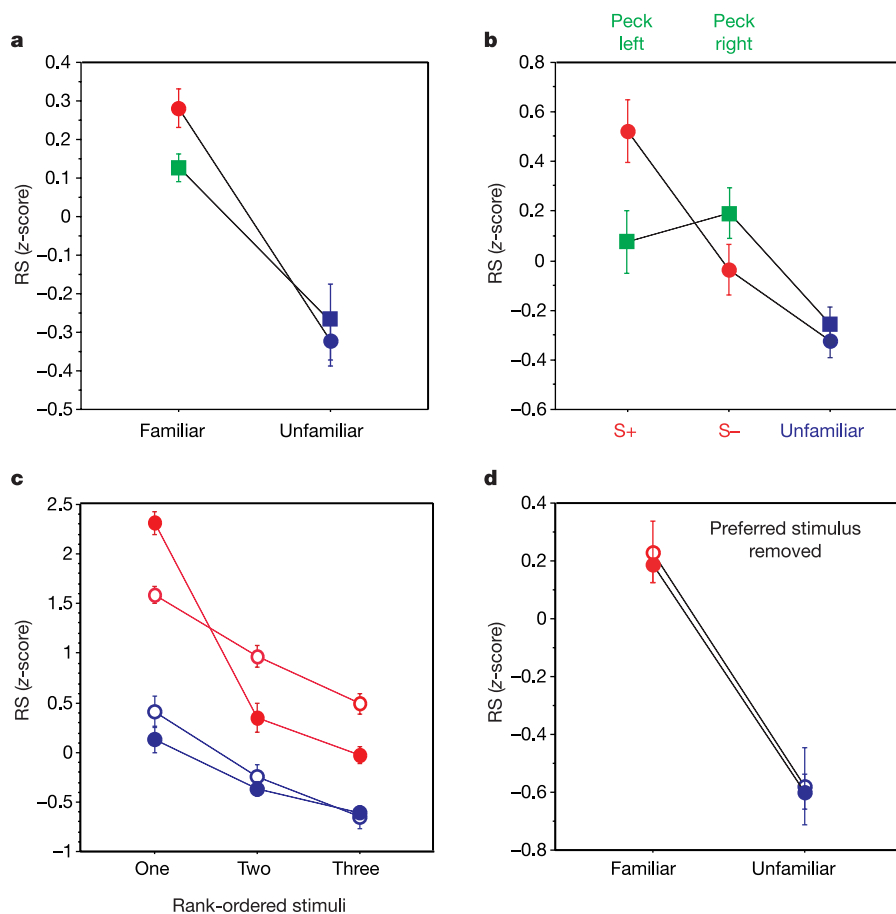


Figure 2 cmHV response strengths. **a**, RS z-scores (see Methods) for familiar (red and green) and unfamiliar (blue) songs, split by training regime (two-alternative choice, squares; go/no-go, circles). **b**, RS z-scores as in **a** but with the two sets of training stimuli and their accompanying responses shown separately (two-alternative choice, green; go/no-go, red). The differences between all three classes for the go/no-go regime were significant (see text). **c**, Rank-ordered RS z-scores for the three most potent familiar (red) and unfamiliar (blue) stimuli for song-selective (filled symbols) and non-selective (open symbols). The interaction between stimulus rank-order and response selectivity

among the familiar songs (red symbols) is significant ($F_{(2,72)} = 23.16$, $P < 0.0001$) and shows the strong bias in the song-selective cells for a single stimulus. The difference between the red and blue curves shows the population-level bias for familiar songs. **d**, RS z-scores for song-selective (filled symbols) and non-selective (open symbols) neurons, with the response to the preferred stimulus (the stimulus that elicited the strongest response) removed from the analysis. The differences are highly significant (see main text). All values are reported as mean $\pm$ s.e.m.

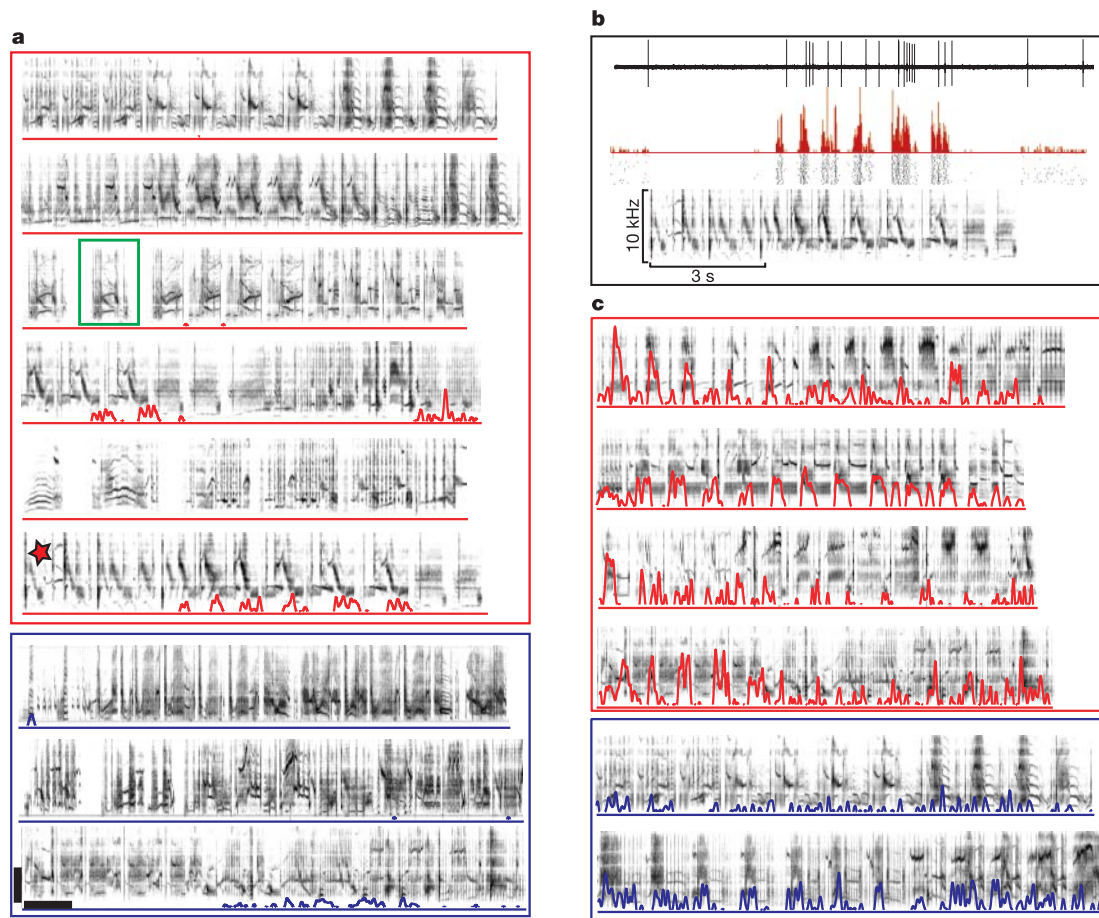


Figure 3 Neuronal responses in the cmHV. **a**, Response of a selective cmHV unit to nine different song stimuli. Familiar songs are outlined in red, unfamiliar songs in blue. The peri-stimulus time histogram (PSTH) of the response is superimposed over the sonogram of each stimulus. The selectivity index, SI, for this cell was the closest to the mean for selective cells (0.497). An example of a single motif is outlined in green. **b**, Detailed view of the response of the unit in **a** to the song stimulus denoted by the red star. Traces from top to bottom show the raw spike waveform for a single stimulus presentation (showing

excellent single-unit isolation), the PSTH, spike raster plots for several stimulus presentations (showing the reliability of the response), and the stimulus sonogram. The response strength, RS, for this song was 8.93. **c**, Example of another selective cmHV neuron responding to six different song stimuli, organized as in **a**. The SI for this cell was 0.259, near the lower limit for selective cells. Horizontal and vertical scale bars show 5 kHz and 1 s, respectively, for **a** and **c**.

mediates a variety of behaviours in both agonistic and antagonistic contexts, different forms of learning under more natural conditions may act on the recognition system to produce a variety of top-down effects on auditory response properties. The presence of these influences on representational plasticity complicates the search for neural correlates in the absence of well-controlled behaviour, and may account for the few reports of such neurons in sensory systems.

Even with the use of extensive stimulus repertoires that animals were trained to recognize, we estimate that less than 50% of the neurons responded to our stimuli (see Methods). The neurons unresponsive to our stimuli may not have been auditory, or may have responded to other songs that these wild-caught birds had previously learned. Given that we find almost no cells selectively tuned to the motifs in unfamiliar songs, the data argue that subpopulations of cmHV neurons are selected from a pool composed primarily of neurons that have already been shaped by the animal's previous experiences with conspecific songs, rather than from a large pool of non-selective neurons. cmHV neurons were more selective than is observed in field L (analogue of the primary auditory cortex) of starlings, where cells commonly respond to numerous novel conspecific vocalizations¹³. Response biases for species-specific vocalizations have been reported in field L of birds^{13,14}, and in mammalian primary and secondary auditory

cortex^{15,16}. Experience-dependent representational plasticity, reported in a variety of animals and sensory systems¹⁷, is typified by shifts in the topography of primary cortical receptive fields^{18–21}. In the auditory system, this plasticity leads to over-representation of the spectral and temporal properties of the stimulus²². Thus, in a hierarchical scheme of sensory processing, plasticity at primary levels should influence higher-order regions, such as cmHV, so that the neuronal response properties and organization are expressly determined by an animal's unique experience on behaviourally relevant tasks. These higher-order regions could contribute to species-specific vocal recognition, and could influence motor tasks such as counter-singing and vocal learning.

Neural correlates of learned object recognition have also been reported in the extrastriate visual cortex, where cells are broadly tuned^{23,24}, and in the prefrontal cortex²⁵, where selectivity for familiar objects is generally evidenced by a decrease in the numbers of neurons responding to a given object — an effect taken to indicate sharpening of the tuning for such objects²⁶. By contrast, the proportion of cmHV cells selective for familiar songs was much larger than that for unfamiliar songs, and many cells were sharply tuned. For vocal recognition, the predictability imparted by species-specific characteristics of vocalizations, and the constraints imposed by evolutionary history and experience, probably yield a population

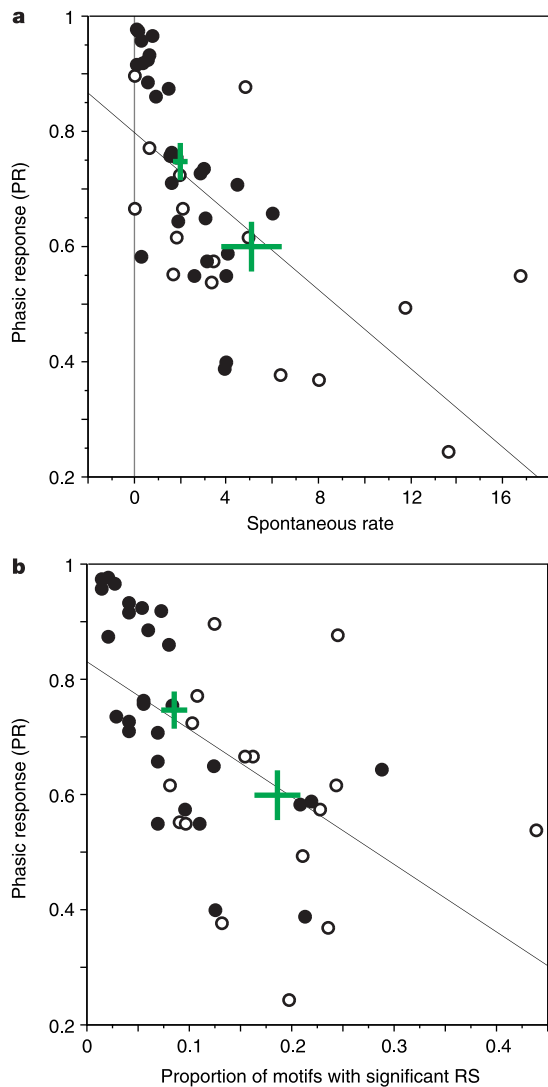


Figure 4 Response scattergrams. **a**, Scatter plot showing the distribution of phasic responses (PR) across the population of selective (filled symbols) and non-selective (open symbols) cmHV neurons, as a function of spontaneous firing rate (spikes s^{-1}). The means ($\pm$ s.e.m.) for each class are shown as the green crosses. The line shows the significant linear regression (Fisher's t to z , $r = -0.65$, $P < 0.0001$). The PR is a normalized measure of the tendency of a cell's discharges to occur in bursts, quantified using the stimulus-driven interspike interval such that the PR for a maximally tonic response is zero, and that for maximally phasic response is 1. The PR of cells depicted in Fig. 3a, c were 0.86 and 0.40, respectively. **b**, As in **a** but with PR plotted as a function of the proportion of motifs that elicited RS values significantly above chance (see Methods). The line shows the significant linear regression (Fisher's t to z , $r = -0.57$, $P < 0.0001$).

of neurons predisposed to represent those vocalizations². We argue that the response properties of cmHV neurons are elaborated continuously towards new functional representations depending on the specific songs and behavioural contingencies an animal encounters. As in juvenile song learning²⁷, the rules by which functional representations arise from biased populations may be complex, and are unlikely to be well predicted by simple spectro-temporal similarities between the target songs and the initial representations of cmHV neurons. The learned representation of auditory objects provides starlings, and perhaps other higher vertebrates, with an efficient mechanism for recognizing a wide and changing array of behaviourally important natural stimuli. □

Methods

Stimuli

Training and testing song stimuli were made from recordings of three adult male starlings, from which we drew three sets of song samples. Each set contained five samples of continuous song from one male (mean sample length, 9.72 ± 0.18 s). Two sets served as operant-training stimuli; the third set served as unfamiliar stimuli during electrophysiological testing. Assignment of each set as either 'training' or 'unfamiliar', and numbers of songs used from each set were counterbalanced across subjects. Subjects were naive to all of the song stimuli before operant training. No motifs were shared between sets, and very few or none within a set. We also presented synthetic noise and frequency-modulated stimuli.

Behavioural training

Six male and two female wild-caught European starlings learned to discriminate between two of the three sets of songs using one of two standard operant-conditioning procedures: a two-alternative choice procedure^{3,4,7} or a go/no-go procedure (see Fig. 1b and Supplementary Information). In each procedure, subjects learned ($n = 4$ per procedure) to recognize the songs in the two training sets using different reward contingencies. All correct responses were positively reinforced in the two-alternative choice procedure, whereas correct responses to only half of the training stimuli (S+) were positively reinforced in the go/no-go procedure (see Supplementary Information).

Electrophysiology

Electrophysiological recordings were conducted using standard extracellular techniques, with free-field stimuli presented to restrained animals inside a sound-isolation chamber. The 45 neurons reported here reflect all the cmHV single neurons that showed significant auditory responses to at least one test stimulus, and for which responses to at least five repetitions (mean, 11.47 ± 0.16) of each test stimulus were obtained. No neurons exhibited only suppression of activity. On the basis of the numbers of cells that did not meet the criterion of five-repetitions per stimulus or show significant auditory responses, we estimate, as an upper bound, that 50% of cmHV cells responded to our stimulus set. We observed no systematic differences between recordings from either hemisphere or sex.

Data analysis

For each neuron, we calculated the mean and variance of the firing rate (spikes s^{-1}) over the duration of each song sample and motif, and during spontaneous activity. We calculated the response strength, RS, associated with each stimulus, as the ratio of stimulus-driven spike-rate variance to the spontaneous spike-rate variance, and used RS as the basis for subsequent analyses. To compare RS values across cells (see, for example, Fig. 2) we converted them to z -scores. Unless noted otherwise, we used a repeated-measures ANOVA (analysis of variance with Greenhouse-Geisser correction) to examine differences in RS z -scores between various stimulus classes (after correcting any deviations from normality). Corrected and uncorrected data yielded statistically identical results.

We quantified each cell's stimulus selectivity using the selectivity index, $SI = (RS_{\max} - \overline{RS})^2 / \sum_{i=1}^N (RS_i - \overline{RS})^2$, where $RS_{\max}$ is the maximum response strength associated with a stimulus, $\overline{RS}$ is the mean response strength across all N test stimuli presented to that cell, and RS_i is the response strength associated with the i th test stimulus for that cell. SI ranges from N^{-1} to 1, with maximal selectivity ($SI = 1$) in the case in which a cell responds to only a single test stimulus. We determined the significance threshold for each cell's SI by simulating (1,000 times) the random response of that cell to the same number of test stimuli and repetitions presented during testing, with RS modelled as a gaussian-distributed random variable (mean = 0, variance = 1), or by the real distribution of RS scores (see Supplementary Information). From the simulated RS data, we compiled a distribution of SI scores for each cell and used the value at the 95th percentile of the simulated SI distribution as the threshold for calling a cell 'selective'.

We compared the proportions of cells selective for familiar songs, unfamiliar songs and artificial stimuli using the chi-squared test. The expected proportions of cells selective for a given stimulus type were equal to the means of the proportions of familiar and unfamiliar stimuli presented to each cell. Mean SI for the population was 0.497 ± 0.033 (mean $\pm$ s.e.m.). Similar analyses using different selectivity thresholds, and with SI scores derived from the mean and variance of the spike rate (rather than RS) produced results qualitatively similar to those reported here (see Supplementary Information).

For the wavelet-decomposition analysis, the proportion of significant multiple regressions for the song-selective cells (17 out of 28, 60.1%) was significantly larger than that for the non-selective cells (3 out of 16, 18.8%; chi-squared test, $P < 0.01$). Identical effects are observed using R^2 values adjusted for the large number of regressors in each model.

Complete descriptions of the stimuli, training, methods and data analyses are provided in the Supplementary Information.

Received 21 January; accepted 22 April 2003; doi:10.1038/nature01731.

1. Kroodsma, D. E. & Miller, E. H. (eds) *The Design of Animal Communication* (MIT Press, Cambridge, Massachusetts, 1999).
2. Gentner, T. Q. & Margoliash, D. in *Acoustic Communication* (eds Simmons, A. M., Popper, A. N. & Fay, R. R.) 324–386 (Springer, New York, 2002).
3. Gentner, T. Q., Hulse, S. H., Bentley, G. E. & Ball, G. F. Individual vocal recognition and the effect of partial lesions to HVC on discrimination, learning, and categorization of conspecific song in adult songbirds. *J. Neurobiol.* **42**, 117–133 (2000).
4. Gentner, T. Q. & Hulse, S. H. Perceptual classification based on the component structure of song in European starlings. *J. Acoust. Soc. Am.* **107**, 3369–3381 (2000).
5. Kroodsma, D. E. & Miller, E. H. (eds) *Ecology and Evolution of Acoustic Communication in Birds* (Cornell Univ. Press, Ithaca, 1996).
6. Eens, M., Pinxten, R. & Verheyen, R. F. Temporal and sequential organization of song bouts in the

- starling. *Ardea* 77, 75–86 (1989).
- Gentner, T. Q. & Hulse, S. H. Perceptual mechanisms for individual vocal recognition in European starlings. *Sturnus vulgaris*. *Anim. Behav.* 56, 579–594 (1998).
 - Vates, G. E., Broome, B. M., Mello, C. V. & Nottebohm, F. Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches. *J. Comp. Neurol.* 366, 613–642 (1996).
 - Müller, C. M. & Leppelsack, H. J. Feature-extraction and tonotopic organization in the avian auditory forebrain. *Exp. Brain Res.* 59, 587–599 (1985).
 - Sen, K., Theunissen, F. E. & Doupe, A. J. Feature analysis of natural sounds in the songbird auditory forebrain. *J. Neurophysiol.* 86, 1445–1458 (2001).
 - Margoliash, D. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J. Neurosci.* 3, 1039–1057 (1983).
 - Margoliash, D. Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *J. Neurosci.* 6, 1643–1661 (1986).
 - Leppelsack, H. J. & Vogt, M. Responses of auditory neurons in forebrain of a songbird to stimulation with species-specific sounds. *J. Comp. Physiol.* 107, 263–274 (1976).
 - Scheich, H., Langner, G. & Bonke, D. Responsiveness of units in the auditory neostriatum of the guinea fowl (*numida-meleagris*) to species-specific calls and synthetic stimuli. 2: discrimination of iambus-like calls. *J. Comp. Physiol.* 132, 257–276 (1979).
 - Wang, X. & Kadia, S. C. Differential representation of species-specific primate vocalizations in the auditory cortices of marmoset and cat. *J. Neurophysiol.* 86, 2616–2620 (2001).
 - Rauschecker, J. P., Tian, B. & Hauser, M. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268, 111–114 (1995).
 - Gilbert, C. D., Sigman, M. & Crist, R. E. The neural basis of perceptual learning. *Neuron* 31, 681–697 (2001).
 - Bakin, J. S. & Weinberger, N. M. Classical conditioning induces CS-specific receptive field plasticity in the auditory cortex of the guinea pig. *Brain Res.* 536, 271–286 (1990).
 - Recanzone, G. H., Schreiner, C. E. & Merzenich, M. M. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J. Neurosci.* 13, 87–103 (1993).
 - Kilgard, M. P. & Merzenich, M. M. Order-sensitive plasticity in adult primary auditory cortex. *Proc. Natl Acad. Sci. USA* 99, 3205–3209 (2002).
 - Kay, L. M. & Laurent, G. Odor- and context-dependent modulation of mitral cell activity in behaving rats. *Nature Neurosci.* 2, 1003–1009 (1999).
 - Kilgard, M. P. & Merzenich, M. M. Plasticity of temporal information processing in the primary auditory cortex. *Nature Neurosci.* 1, 727–731 (1998).
 - Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5, 552–563 (1995).
 - Kobatake, E., Wang, G. & Tanaka, K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophysiol.* 80, 324–330 (1998).
 - Rainer, G., Asaad, W. F. & Miller, E. K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* 393, 577–579 (1998).
 - Rainer, G. & Miller, E. K. Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron* 27, 179–189 (2000).
 - Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* 291, 2564–2569 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank F. E. Theunissen for generously providing software and assistance for calculating STRFs, M. Konishi and L. M. Kay for valuable critiques of the manuscript, Z. Chi and S. Shea for helpful discussions, and D. Baleckaitis for histology. This work was supported by grants from the National Institutes of Health to T.Q.G. and D.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.Q.G. (t-gentner@uchicago.edu).

Neural correlates of implied motion

Bart Krekelberg¹, Sabine Dannenberg², Klaus-Peter Hoffmann², Frank Bremmer^{2,3} & John Ross⁴

¹Vision Center Laboratory, The Salk Institute, La Jolla, California 92037, USA

²Department of Neurobiology, Ruhr University, D-44780 Bochum, Germany

³Cognitive Neuroscience Laboratory, Department of Neurophysics, Philipps University Marburg, D-35032 Marburg, Germany

⁴School of Psychology, The University of Western Australia, Nedlands, WA 6907, Australia

Current views of the visual system assume that the primate brain analyses form and motion along largely independent pathways¹; they provide no insight into why form is sometimes interpreted as motion. In a series of psychophysical and electrophysiological experiments in humans and macaques, here we show that some

form information is processed in the prototypical motion areas of the superior temporal sulcus (STS). First, we show that STS cells respond to dynamic Glass patterns², which contain no coherent motion but suggest a path of motion³. Second, we show that when motion signals conflict with form signals suggesting a different path of motion, both humans and monkeys perceive motion in a compromised direction. This compromise also has a correlate in the responses of STS cells, which alter their direction preferences in the presence of conflicting implied motion information. We conclude that cells in the prototypical motion areas in the dorsal visual cortex process form that implies motion. Estimating motion by combining motion cues with form cues may be a strategy to deal with the complexities of motion perception in our natural environment.

Many studies have shown that the ventral part of the primate brain specializes in the processing of complex visual form, whereas the dorsal part of the brain specializes in the analysis of visual motion¹. Some, however, have questioned this strict division of labour. Geisler⁴, for instance, showed that humans use traces of motion, which he termed motion streaks, to improve their detection of moving objects. Cells in the primary visual cortex could underlie this ability^{5,6}. Kourtzi and Kanwisher⁷ showed that static images that imply motion (such as a basketball player about to throw a ball) activate areas in the human brain that process motion information (hMT+). Finally, Ross *et al.*³ demonstrated that humans can perceive form itself as motion: subjects perceived coherent motion in sequences of rotational Glass patterns² that contain no coherent motion.

Glass patterns consist of a collection of randomly placed pairs of dots, all oriented along a common path (Fig. 1a). This path corresponds to the perceived path of motion. We refer to this type of motion, which is ambiguous in direction, as implied motion; it is implied by the form of the global pattern, not carried by directed motion signals (see Supplementary Information A). The perception of coherent motion in Glass patterns depends critically on the separation (called the Glass shift) between the elements of the dot pairs. For small and large shifts, the sequence of patterns is perceived as random noise; only for an intermediate range does the percept of motion arise (see Fig. 2 in Ross *et al.*³). By aligning dot pairs along different paths, various types of global illusory motion can be created: translations, rotations, and expansions and contractions. The global organization in Glass patterns is particularly strong for rotations^{3,8} and expansions⁹. Cells in visual area V1 in macaques have recently been shown to respond selectively to the weak locally oriented features¹⁰, allowing them to pass on this information to extrastriate areas for the analysis of global form. These properties make Glass patterns particularly suitable for our experiments. Nevertheless, we believe that they are merely instances of a general class of stimuli that provide motion information by means of their form⁹.

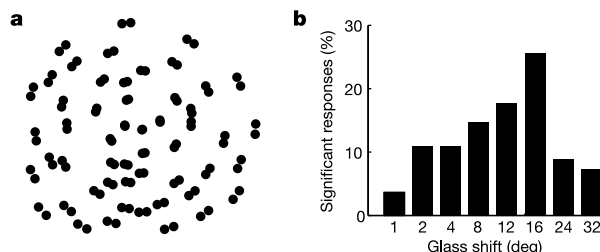


Figure 1 Optimal Glass patterns activate the STS population. **a**, A rotational Glass pattern. Presenting many randomly chosen patterns like this, one after another, leads to the impression of rotational motion³. **b**, The histogram shows how likely it was that the presentation of a Glass pattern with a given shift evoked a response in any of the STS cells that was significantly larger than the random-noise control condition.

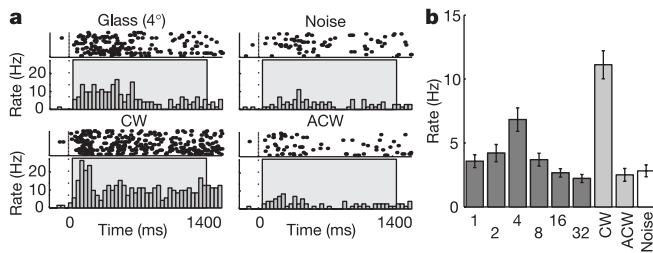


Figure 2 Responses to implied motion in a single cell. **a**, Raster and peri-stimulus-time histogram for responses to a Glass pattern with a 4° shift, a noise pattern, and clockwise (CW) and anticlockwise (ACW) rotation. **b**, Average firing rates in the stimulation period indicated by the grey area in **a**. The numbers on the ordinate refer to the Glass shift in degrees.

We have found a neural correlate of the global implied motion percept in the dorsal stream of the macaque visual cortex. While the monkeys fixated the centre of a screen, we recorded from cells in the medial (MT) and medial superior (MST) temporal areas of the superior temporal sulcus. We presented sequences of rotational Glass patterns (see Methods) with a varying amount of Glass shift. Humans perceived the patterns with the smallest and largest shift as random noise; only the patterns in the centre of the tested range reliably led to a coherent rotational motion percept, albeit with an ambiguous direction. Figure 2 shows a cell that fired maximally when the Glass shift was such that humans perceived coherent rotation. For small or very large Glass shifts, however, the firing of the cell was indistinguishable from its response to random noise. Of the 109 cells tested in this paradigm, 18 (17%) showed such a tuning for Glass shift. Figure 2 also shows that this cell's response to Glass patterns was more variable over time. This was typical. Because the direction of motion in these patterns is ambiguously clockwise and anticlockwise, one may in fact have predicted that the neurons should show clockwise responses on some trials and anticlockwise responses on other trials. We could show that this was indeed often the case (see Supplementary Information B).

Next, we determined whether the STS population as a whole was sensitive to those Glass patterns that we perceived as coherent motion. Figure 1b shows the probability that a particular Glass shift evoked a mean response larger than the response to random noise (analysis of variance on ranks, followed by a post-hoc test, 109 cells). This confirms that Glass patterns outside the central range tended to evoke a response that was indistinguishable from the response to random noise. In the centre of the Glass shift range, where we perceived coherent motion, more responses were significantly larger than the response to random noise.

To further demonstrate the equivalence of implied and real

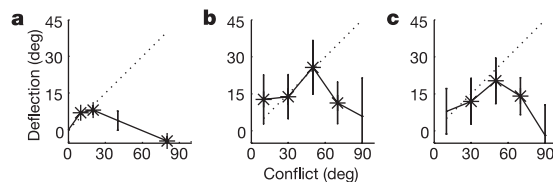


Figure 3 Interactions between real and implied motion. **a**, Human perception. The deflection of direction—averaged over three observers—is shown as a function of the conflict angle. **b**, The average STS population deflection as a function of the conflict between the orientation of the Glass pattern and the PD of the cell. **c**, The deflection as a function of the conflict between the axis of real motion and the PD. In all three panels, a positive deflection indicates a change towards the implied motion, asterisks indicate statistical significance, the dotted line represents the predicted deflection if real and implied motion had equal influence, and error bars represent 95% confidence intervals.

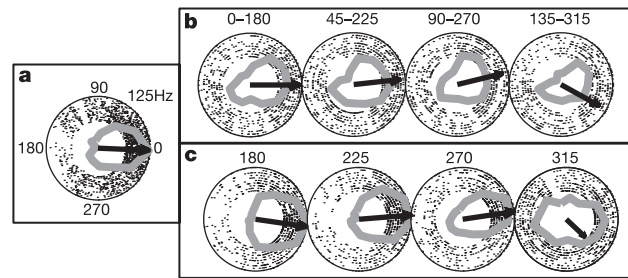


Figure 4 Interactions between real and implied motion in a single cell. **a**, Polar direction tuning plot for a random dot pattern on a circular pathway. Dots represent single spikes recorded while the pattern moved in that direction, corrected for the latency of the cell. The solid grey curve represents the mean firing rate (the full radius represents 125 Hz in all polar plots), the arrow the preferred direction. **b**, Direction tuning for a combination of circular pathway motion with a Glass pattern oriented along the axis indicated above the plots. **c**, Direction tuning for circular pathway motion combined with translational motion in the direction indicated above the plots.

motion, we determined whether a cell's preference for expanding or rotating flow patterns predicted its preference for expansion or rotational Glass patterns. The trend ($P = 0.055$) that 20 out of 32 cells thus tested preferred the same type of motion in the real motion domain as in the implied motion domain suggests that implied motion activated the appropriate cells in the STS.

If indeed the same neurons process implied and real motion cues, one would expect these cues to interact perceptually. In other words, it should be possible to change the perceived direction of real motion by giving conflicting implied motion information. We first tested this in a psychophysical experiment in humans.

Subjects judged the direction of global motion in displays containing both real-motion and translational Glass patterns. Figure 3a shows by how much the judged direction of global motion was deflected away from the true direction and attracted towards the direction implied by the Glass patterns. We call the difference in angle between the implied motion and the true direction of motion the conflict angle. When the conflict angle was small—less than 20°—the deflection was almost half the angular difference. This indicates that, under these circumstances, the implied motion and the real motion had almost equal influence on the judged direction of motion. As the conflict angle widened, the influence of pair orientation on judged motion direction weakened. Near 90° the Glass pattern had only a small influence—repulsive, not attractive—on the judged direction of motion. An even stronger influence of form on motion was found for spiral Glass patterns (Supplementary Information C).

Next, we investigated the interaction between real and implied motion in STS cells. First, we determined directional tuning curves for each cell by showing a field of random dots on a circular motion

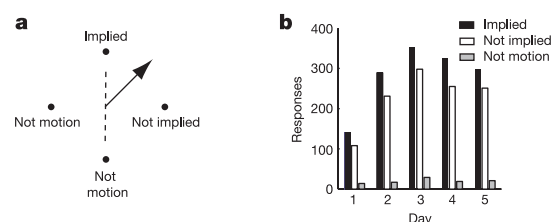


Figure 5 Monkey psychophysics. **a**, The arrow indicates the direction of real motion, the dashed line the orientation of the Glass pattern. The dots are saccade targets that the monkey used to indicate his perceived direction of motion. In this example (45° real motion with a vertical Glass pattern), the monkey's decisions can be interpreted as indicated by the labels. **b**, Results, averaged over all directions of motion, on five consecutive days.

trajectory in the fronto-parallel plane. This gave us a preferred direction (PD) of motion for the neuron^{11,12}. Second, we mapped the directional tuning again, but now while a translational Glass pattern oriented along one of the cardinal axes was shown transparently. Of the 62 visually responsive cells recorded in this paradigm, the Glass patterns significantly affected the preferred direction of 16 (26%) cells (one-way analysis of variance on ranks, $P < 0.05$). Third, we mapped the directional tuning again, but now while a real translating flow field was presented transparently. The axis of motion of the real translating flow field significantly affected the preferred direction of 37 neurons (60%). Assuming that cells represent the direction of motion for which they respond best, our psychophysical findings predicted that Glass patterns that were orthogonal to a cell's PD (large conflict angle) should have only a small influence on the PD. For Glass patterns that were nearly but not quite parallel to the PD of a cell (small conflict angle), we expected the PD to shift towards the implied motion cue in the Glass pattern. Figure 4 shows an example of a cell that fitted this prediction. The cell preferred rightwards motion (Fig. 4a). As expected, the combination of multiple motion signals (Fig. 4b, c) resulted in a reduced firing rate^{13,14} and broader tuning curves. Nevertheless, when direction tuning was assessed in the presence of a 135–315° Glass pattern or a 315° real motion pattern, this cell's preferred direction shifted consistently towards 315°. For this cell, the effect of all other directions was smaller. Moreover, the effects of real and implied motion were similar. For individual cells, the changes in PD were often small and variable; in Fig. 4, only the change in PD with a 135–315° Glass pattern is significantly different from zero. At the population level, however, the effects were clearer. Figure 3b, c shows population averages of the deflection as a function of the conflict angle. Figure 3b demonstrates that Glass patterns that were close to the preferred direction of the cell changed the preferred direction for the transparent stimulus. For small conflicts between the implied motion and real motion signals, the influence of implied motion was as strong as that of real motion; the represented direction of motion was the vector average. For larger conflicts, however, the cells demonstrated winner-take-all behaviour: they ignored the implied motion. Figure 3c shows the same calculations for the interaction of two transparently shown directions of real motion. Here, too, small conflicts lead to deflections of the preferred direction, whereas large conflicts show winner-take-all behaviour. In the human psychophysics we found that large conflicts lead to a small repulsive effect; even though there is a small negative deflection in the motion–motion interactions in the STS (Fig. 3c), this effect was not statistically significant. These data not only show a neural correlate of the psychophysical experiment in Fig. 3a, they also show that the influence of implied motion on real motion (Fig. 3b) is very similar to the influence of real motion on real motion (Fig. 3c). This gives further support to our claim that the visual system's processing of implied motion is very similar to that of real motion.

The interaction between real and implied motion allowed us to come full circle and test whether monkeys perceived implied motion. We trained monkey T on a four-alternative forced-choice direction-discrimination task (Fig. 5a). During the training phase, random dots moved in one of eight directions for 1 second. Then, four dots appeared and the monkey reported his perceived direction of motion by making a saccade to one of the dots. Performance reached ~80% correct. We then tested how a simultaneously presented translational Glass pattern influenced the decisions. For example, we presented 45° motion transparently with a vertical Glass pattern and showed saccade targets only for the cardinal directions. Our human psychophysical data predicted that the vertical Glass pattern should cause an upward motion bias in the 45° motion stimulus. Hence, a decision for the upward motion direction would support the implied motion hypothesis (an 'implied motion' decision), a decision for the rightward motion

direction, however, counted against the hypothesis ('not-implied'). A decision for either left or downward motion is evidence that the monkey was not paying attention to the motion at all ('not-motion'). In ambiguous motion trials, the reward was random at a rate proportional to the monkey's performance on interleaved trials in which only unambiguous real motion was shown. We performed this experiment on five consecutive days and consistently found a significant ($P < 0.05$) implied motion bias for all motion directions. On average, the monkey was ~11% more likely to make a decision in the implied motion direction than orthogonal to it (Fig. 5b). The 'not-motion' decisions were rare. Hence, even on the randomly rewarded trials, the monkey attempted to perform a motion direction task. The bias towards the implied motion direction confirms that form biases motion percepts in the macaque as it does in humans.

The responses we find may be related to the orientation-selective responses in MT that have been described before^{15–17}. Given that MT inactivation impairs direction but not orientation discrimination^{18,19}, however, the relevance in MT of orientation tuning *per se* is somewhat mysterious. The preference of some cells in MT for motion parallel to their preferred orientation is commonly interpreted in terms of the intersection of constraints model¹⁶. But, as Geisler *et al.* have pointed out, it could also be interpreted as sensitivity to the motion implied by streaks. Motion streaks provide the best motion information at high speeds, where estimates of real motion become less reliable; hence these motion mechanisms may in fact be complementary^{4,5}.

The unity of perception requires that sooner or later information on form be combined with information on motion. Our results show that at least some combination occurs not after the dorsal stream has completed its processing of motion signals, but while it is doing so. Taken together with the data of Kourtzi and Kanwisher⁷, it thus appears that many hints of motion, like the oriented pairs in our Glass patterns, modulate cells in the motion areas of human and monkey brains. This may reflect a strategy that combines form with motion information to survive in an environment in which combinations of object motion and self motion, occlusion, and transparency complicate the tasks our visual system has to solve. □

Methods

Human psychophysics

Stimuli were displayed on a Hitachi Accuvue 4821 monitor (120 Hz, 800 pixels × 600 pixels) with a Cambridge Research Systems VSG2/4 graphics card. Noisy translational Glass patterns, composed of white dots (diameter 5'; 93 cd m⁻²) were displayed on a circular grey disc (diameter 14.5'; 21.5 cd m⁻²). The first pattern in each sequence of 10 was produced by randomly positioning the first dot of each pair, then placing its partner 20' away, either at the signal angle for signal pairs, or at a random angle for noise pairs. Subsequent patterns were made by moving the appropriate proportion of pairs and randomly replacing the remainder. Motion step size was 20' and new patterns were shown with 12 Hz, giving a dot speed of 4° s⁻¹.

Procedure

Motion direction was 15°, 30°, 45° or 60° from horizontal, and the orientation of Glass pairs deviated from each motion direction (positively or negatively) by 0°, 10°, 20°, 40° or 80°; 25% of pairs moved in the signal direction and 50% were at the signal orientation. Subjects reported the direction of global motion verbally or by mouse-click. Results were averaged over four judgments. Two subjects were unaware of the purpose of the experiment; one was an author.

Monkey physiology

We used two male macaques (C and T). Animal treatment was in accordance with European Community (86/609/ECC) and NIH guidelines. We recorded 168 cells in the medial temporal and medial superior temporal areas in two hemispheres. Not every paradigm could be recorded for every cell. Identification of areas on the basis of MR images, chamber position, electrode depth, the large fraction of directionally selective cells with small receptive fields (MT) or a preference for flow fields and large receptive fields (MST) led to the estimate that one-third of cells were in MT and two-thirds were in MST. We found no clear differences in the responses to the Glass patterns and therefore treated them as a single population, referred to as STS. The monkey fixated a small (<0.5°) central red dot. Fixation within 2° was monitored with an eye-coil system (500 Hz, accuracy ~1'). A frame in a Glass sequence consisted of 50 randomly positioned pairs of dots (diameter ~0.25°) in the central 30° (monkey C) or 15° (monkey T). In rotational Glass patterns, these were oriented along concentric circles around the fixation spot. In translations, the pairs were oriented horizontally (0°) or at 45°, 90° or 135°. In expansions, the pairs were

oriented along radii centred at fixation. Each frame was constructed from a new set of pairs. Initially, frames were refreshed at 50 Hz, later in the series of experiments at 12 Hz. In the rotations, we varied the separation between the elements of a pair (the Glass shift) initially over 1°, 2°, 4°, 8°, 16° and 32° of rotation; later we added 12° and 24°. To correct for the unequal number of presentations, the probabilities in Fig. 1 were calculated as a fraction of the number of times that a shift was used. To stimulate MT cells adequately, we used larger dots than in the psychophysics; the range of shifts was selected such that human observers judged only the centre of the range to be moving. In the translations, the shift was 1°, a value that led to a consistent motion percept in human observers.

In the 'noise' condition, 100 random dots were randomly repositioned at 50 Hz or 12 Hz. In the clockwise and anticlockwise conditions, 100 random dots rotated coherently. The rotation speeds (angular velocity 240° s⁻¹) matched the perceived speed in the Glass patterns. In any experiment, the dot density, extent and luminance were the same for all patterns.

To determine direction tuning, 100 random dots moved on a circular fronto-parallel pathway. This generated motion in all directions and a tuning curve could quickly be generated¹¹. Latency was defined as the first 40-ms bin in which an onset response significantly ($P < 0.05$) exceeded the rate in the 200 ms before stimulus onset. To determine the preferred direction (PD), we corrected for the latency, and determined the circular pathway response in 40-ms bins. The PD was defined as the centroid of this activity. The Rayleigh test²⁰ gave a measure of its significance.

In the interaction paradigms, we combined the circular pathway with Glass patterns or real translations. We called the angle between the PD of the cell and the orientation of the Glass pattern (or the axis of motion for real translation) the conflict angle. Using the axis of motion allowed us to compare the effect of real motion with that of the directionally ambiguous Glass patterns. The change in PD due to the additional pattern was called the deflection. To determine the population deflection, we binned the deflections with respect to the conflict angle (bin width 20°) and averaged over all cells. Its significance was assessed with a Rayleigh test.

Received 13 February; accepted 10 June 2003; doi:10.1038/nature01852.

- Mishkin, M., Ungerleider, L. & Macko, K. Object vision and spatial vision: two central pathways. *Trends Neurosci.* **6**, 414–417 (1983).
- Glass, L. Moiré effect from random dots. *Nature* **223**, 578–580 (1969).
- Ross, J., Badcock, D. R. & Hayes, A. Coherent global motion in the absence of coherent velocity signals. *Curr. Biol.* **10**, 679–682 (2000).
- Geisler, W. S. Motion streaks provide a spatial code for motion direction. *Nature* **400**, 65–69 (1999).
- Geisler, W. S., Albrecht, D. G., Crane, A. M. & Stern, L. Motion direction signals in the primary visual cortex of cat and monkey. *Vis. Neurosci.* **18**, 501–516 (2001).
- Jancke, D. Orientation formed by a spot's trajectory: A two-dimensional population approach in primary visual cortex. *J. Neurosci.* **20**, RC86 (1–6) (2000).
- Kourtzi, Z. & Kanwisher, N. Activation in human MT/MST by static images with implied motion. *J. Cogn. Neurosci.* **12**, 48–55 (2000).
- Wilson, H. R., Wilkinson, F. & Asaad, W. Concentric orientation summation in human form vision. *Vision Res.* **37**, 2325–2330 (1997).
- Burr, D. C. & Ross, J. Direct evidence that 'Speedlines' influence motion mechanisms. *J. Neurosci.* **22**, 8661–8664 (2002).
- Smith, M. A., Bair, W. & Movshon, J. A. Signals in macaque striate cortical neurons that support the perception of Glass patterns. *J. Neurosci.* **22**, 8334–8345 (2002).
- Schoppmann, A. & Hoffmann, K.-P. Continuous mapping of direction selectivity in the cat's visual cortex. *Neurosci. Lett.* **2**, 177–181 (1976).
- Bremmer, F., Ilg, U. J., Thiele, A., Distler, C. & Hoffmann, K. P. Eye position effects in monkey cortex. I. Visual and pursuit-related activity in extrastriate areas MT and MST. *J. Neurophysiol.* **77**, 944–961 (1997).
- Snowden, R. J., Treue, S., Erickson, R. G. & Andersen, R. A. The response of area MT and V1 neurons to transparent motion. *J. Neurosci.* **11**, 2768–2785 (1991).
- Qian, N. & Andersen, R. A. Transparent motion perception as detection of unbalanced motion signals. II. Physiology. *J. Neurosci.* **14**, 7367–7380 (1994).
- Maunsell, J. H. & Van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *J. Neurophysiol.* **49**, 1127–1147 (1983).
- Albright, T. D. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* **52**, 1106–1130 (1984).
- Pack, C. & Born, R. T. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature* **409**, 1040–1042 (2001).
- Newsome, W. T. & Pare, E. B. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* **8**, 2201–2211 (1988).
- Rudolph, K. & Pasternak, T. Transient and permanent deficits in motion perception after lesions of cortical areas MT and MST in the macaque monkey. *Cereb. Cortex* **9**, 90–100 (1999).
- Batschelet, E. *Circular Statistics in Biology* (Academic, London, 1981).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Constanza, D. Diep, L. Abavare and M. Bronzel for technical assistance, C. Distler for the surgeries, and G. Stoner and T. Albright for comments and discussions. The Human Frontiers Science Program and the Australian Research Council supported this work financially.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.K. (bart@salk.edu).

Induction of dendritic spines by an extracellular domain of AMPA receptor subunit GluR2

Maria Passafaro^{1,2}, Terunaga Nakagawa³, Carlo Sala² & Morgan Sheng³

¹DTI Dulbecco Telethon Institute,

²CNR Institute of Neuroscience, Cellular and Molecular Pharmacology, Department of Pharmacology, University of Milan, 20129 Italy

³Picower Center for Learning and Memory, RIKEN-MIT Neuroscience Research Center, Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Synaptic transmission from excitatory nerve cells in the mammalian brain is largely mediated by AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid)-type glutamate receptors located at the surface of dendritic spines. The abundance of postsynaptic AMPA receptors correlates with the size of the synapse and the dimensions of the dendritic spine head^{1–4}. Moreover, long-term potentiation is associated with the formation of dendritic spines as well as synaptic delivery of AMPA receptors^{5–8}. The molecular mechanisms that coordinate AMPA receptor delivery and spine morphogenesis are unknown. Here we show that overexpression of the glutamate receptor 2 (GluR2) subunit of AMPA receptors increases spine size and density in hippocampal neurons, and more remarkably, induces spine formation in GABA-releasing interneurons that normally lack spines. The extracellular N-terminal domain (NTD) of GluR2 is responsible for this effect, and heterologous fusion proteins of the NTD of GluR2 inhibit spine morphogenesis. We propose that the NTD of GluR2 functions at the cell surface as part of a receptor–ligand interaction that is important for spine growth and/or stability.

In mature cultured hippocampal neurons (22 days *in vitro*; DIV22), which normally exhibit mushroom-like spines, overexpression of GluR2 increased the length of spines, the width of spine heads and the density of spines (Figs 1 and 2d; see also Supplementary Table 1). In younger neurons (DIV11), GluR2 overexpression induced a higher density of filopodia-like protrusions that were longer and wider than those found on control neurons (Fig. 1; see also Supplementary Table 1). GluR1 overexpression had little effect on the size or number of dendritic spines in DIV22 neurons, and did not induce filopodia-like protrusions in DIV11 neurons (Figs 1 and 2d; see also Supplementary Table 1). Neurons transfected with GluR3 showed spines similar to control green fluorescent protein (GFP)-transfected neurons (Supplementary Table 1). Thus the ability to increase spine number and size seems to be specific for GluR2.

The enlarged spines of GluR2-transfected neurons probably represent bona fide postsynaptic compartments, as they were enriched for F-actin, contained clusters of the postsynaptic density (PSD) proteins Shank and PSD-95, and co-localized with the presynaptic protein bassoon. Indeed, spines of GluR2-overexpressing cells showed increased staining intensity for Shank, PSD-95, F-actin and bassoon, consistent with expansion of the PSD and presynaptic growth (Supplementary Fig. 1).

Which domain(s) of GluR2 are required for its spine-promoting effect? GluR2 mutants lacking the last four amino acids (GluR2Δ4) or the entire cytoplasmic tail (GluR2Δ50) also caused increased spine density and spine-head enlargement (Supplementary Fig. 2 and Table 1), even though the GluR2Δ50 mutant showed reduced surface expression (Supplementary Fig. 3). Thus the carboxy-terminal tail of GluR2, which mediates multiple interactions with

cytoplasmic proteins, is not important for the spine-promoting activity of GluR2.

A mutation of a channel pore residue (R607Q) of GluR2, which converts arginine to glutamine at position 607 (thereby allowing calcium permeability of the GluR2 receptor), was still able to increase spine density (Supplementary Fig. 2 and Table 1). GluR2(R607Q) also increased the width and length of spines, but to a lesser extent than wild-type GluR2. These results suggest that the channel properties of GluR2 are not important for its trophic effect on spines.

The spine-promoting activity of GluR2 was abolished by deletion

of the extracellular NTD (residues 60–398; GluR2 Δ NTD), also known as the amino-terminal domain or ATD⁹. Indeed, in neurons overexpressing GluR2 Δ NTD, spine density was reduced by >60%, and spine length and spine-head width were strongly decreased (Fig. 2; see also Supplementary Table 1), suggesting a dominant negative effect of GluR2 Δ NTD on spine morphogenesis. The surface expression of GluR2 Δ NTD after transfection was similar to wild-type GluR2 (Supplementary Fig. 3). Overexpression of the cytoplasmic tail of GluR2 fused to GFP (GFP–R2C), which interferes with synaptic delivery of GluR2 in electrophysiological assays¹⁰, caused a significant reduction in spine-head width but only a nonsignificant decrease in spine length and density (Supplementary Fig. 2 and Table 1).

We analysed a series of chimaeras between GluR2 and GluR1 to verify that the NTD is important for spine induction and growth. Notably, a chimaera in which the NTD of GluR1 is replaced with the NTD of GluR2 (NTDR2/R1) was as effective as wild-type GluR2 in increasing spine density and spine width (Fig. 2; see also Supplementary Table 1). Conversely, a chimaera in which the NTD of GluR2 is replaced with the NTD of GluR1 (NTDR1/R2) had no effect on spine density, and actually caused a significant reduction in spine-head size, consistent with a dominant-interfering effect. These results confirm the NTD as the critical determinant of the spine-promoting activity of GluR2.

To test whether the endogenous GluR2 NTD is involved in spine morphogenesis, we treated hippocampal cultures with soluble fusion protein of the immunoglobulin constant region (Fc) fused to the NTD of GluR2 (Fc–NTDR2) or GluR1 (Fc–NTDR1). Fc–NTDR2, but not Fc–NTDR1 or Fc alone, caused a significant reduction in spine density (Fig. 3b, c; see also Supplementary Table 1). Similar results were obtained when hippocampal neurons were transfected with the Fc–NTD constructs. Specifically Fc–NTDR2 reduced spine density and increased spine length (Fig. 3a, c; see also Supplementary Table 1). The increased spine length in Fc–NTDR2-transfected neurons may stem from a higher proportion of filopodia-like protrusions.

For further corroboration, we inhibited expression of GluR2 protein in hippocampal neurons by small interfering RNA (siRNA; GluR2 immunostaining intensity was decreased to <10% of vector-transfected controls, but GluR1 staining was unaffected (data not shown)). GluR2 ‘knockdown’ caused a significant reduction in spine density and spine width (Fig. 3d–f; see also Supplementary Table 1). These findings support the idea that specific interactions mediated by the NTD of endogenous GluR2 are important for spine morphogenesis.

The effects of GluR2 (see above) were measured in presumptive pyramidal neurons. Notably, GluR2 transfection induced spine formation even in GABA (γ -aminobutyric acid)-releasing interneurons, non-spiny neurons that can be identified in culture by their distinctive dendritic morphology and their cell body staining for glutamic acid decarboxylase (GAD) and/or parvalbumin (Fig. 4). GABA-releasing interneurons transfected with GluR2 showed spine-like dendritic protrusions that were similar in density and shape to normal spiny neurons (Fig. 4; see also Supplementary Table 1), and these protrusions also co-localized with Shank and bassoon puncta (data not shown). GABA-releasing neurons transfected with haemagglutinin (HA)-tagged GluR1 and GFP never displayed spine-like protrusions on their dendrites and were morphologically indistinguishable from control GFP-transfected interneurons (Fig. 4).

Our findings reveal a direct link between AMPA receptors and spine formation and growth. A role for GluR2 in spine morphogenesis could explain the close correlation between AMPA receptor content of synapses and the size and shape of dendritic spines^{1–4}. The spine-promoting activity of GluR2 is highlighted by its ability to induce spines in GABA-releasing interneurons. This result is provocative because these non-spiny neurons typically express low

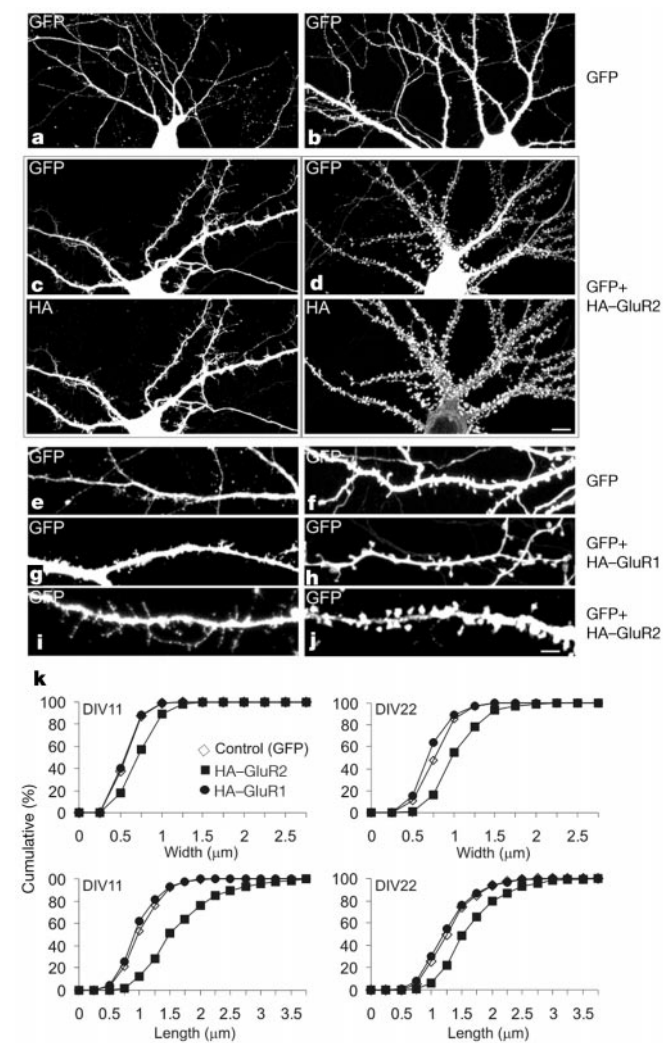


Figure 1 GluR2 promotes increased number and size of spines. Cultured hippocampal neurons were transfected with enhanced GFP (EGFP) alone, or co-transfected with EGFP and HA epitope-tagged GluR1 or GluR2, as indicated. One week after transfection, neurons were stained for GFP and HA at DIV11 (left panels) and DIV22 (right panels). Paired panels in **c** and **d** are double-labelled images of GFP– and HA–GluR2 in the same neuron, as indicated in each panel. In DIV11 neurons, GluR2 promotes filopodia-like protrusions (top panel in **c** and **i**) longer than those found on control GFP-transfected neurons (**a**, **e**). In DIV22 neurons, GluR2 promotes enlargement of spines (compare top panel in **d** with **b**, and **j** with **f**). Overexpression of HA–GluR1 has little effect on dendritic morphology at DIV11 (**g**) or DIV22 (**h**). Scale bars, 10 μ m (low magnification) and 2.5 μ m (high magnification). **k**, Cumulative frequency plots of spine width and spine length in neurons transfected with EGFP, EGFP plus HA–GluR1 or with EGFP plus HA–GluR2 (>1,000 spines and >12 neurons examined for each construct), at DIV11 and DIV22. All morphometric measurements were performed using GFP images.

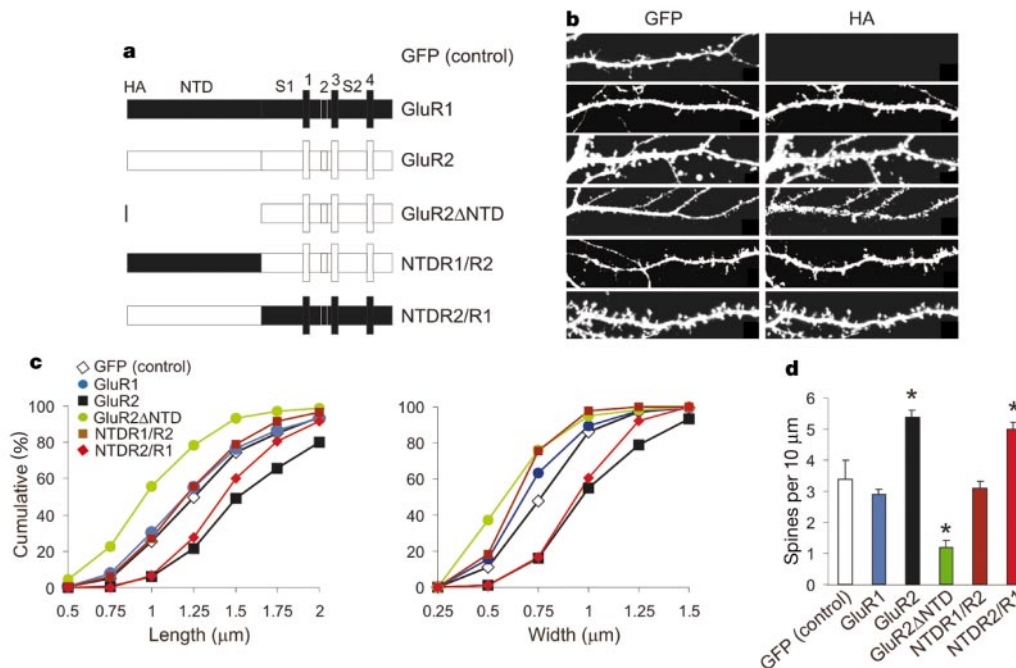


Figure 2 Spine-inducing determinants of GluR2. **a**, Diagram of GluR2ΔNTD and chimeric constructs between GluR1 and GluR2 (all HA-tagged at the N terminus). **b**, Hippocampal neurons were transfected with EGFP alone (control) or EGFP plus various HA–GluR constructs, as indicated on the left. Each pair of images shows the transfected GluR construct stained by HA antibodies (right) and GFP fluorescence to outline dendrite/

spine morphology (left). **c**, Cumulative frequency plots of spine length and spine-head width in neurons transfected as in **b** (>1,000 spines and >12 neurons examined for each construct). **d**, Quantification of spine density (number of spines per 10 μm dendrite; mean ± s.e.m.) in neurons transfected as in **b** (>12 neurons, >140 dendrites examined for each construct). Asterisk, $P < 0.001$.

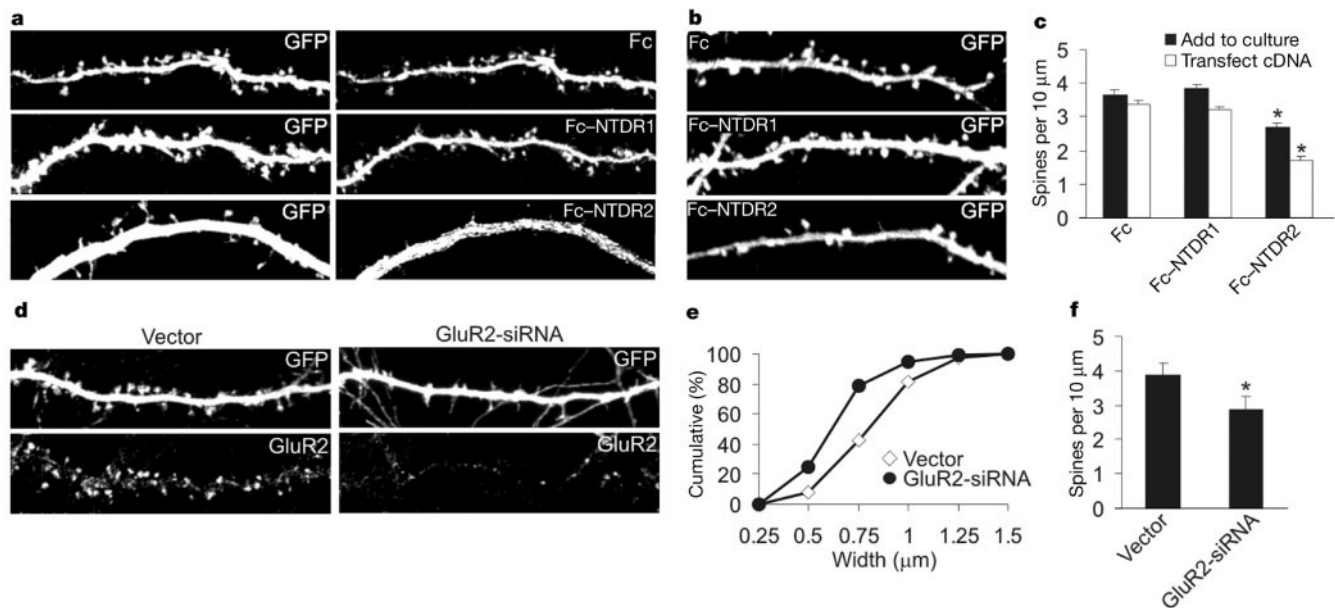


Figure 3 Inhibition of spine morphogenesis by Fc fusion of GluR2 NTD and siRNA knockdown of GluR2. **a**, Hippocampal neurons were co-transfected at DIV14 with EGFP plus the indicated Fc–NTD construct. Each pair of panels shows double labelling for GFP and Fc at DIV22. **b**, Neurons (DIV14) were transfected with EGFP for visualization of morphology and then treated for 6 days (5 μg protein per well, refreshed daily) with soluble immunoglobulin Fc alone, or Fc fusion protein of GluR1 NTD (Fc–NTDR1) or GluR2 NTD (Fc–NTDR2). **c**, Quantification of spine density (number of spines per 10 μm; mean ± s.e.m.) in neurons transfected as in **a** or treated as in **b** (40 neurons examined

for each treatment). Asterisk, $P < 0.01$. **d**, Hippocampal neurons (DIV12) were co-transfected with EGFP plus GluR2-siRNA plasmid construct (right panels) or siRNA vector (left), and double labelled at DIV18 for GFP and for surface GluR2 as indicated in each pair of panels. **e**, Cumulative frequency plots of spine-head width in neurons transfected as in **d** (>1,000 spines and >25 neurons examined for each construct). **f**, Quantification of spine density (number of spines per 10 μm) in neurons transfected as in **d** (25 neurons examined for each treatment). Asterisk, $P < 0.01$.

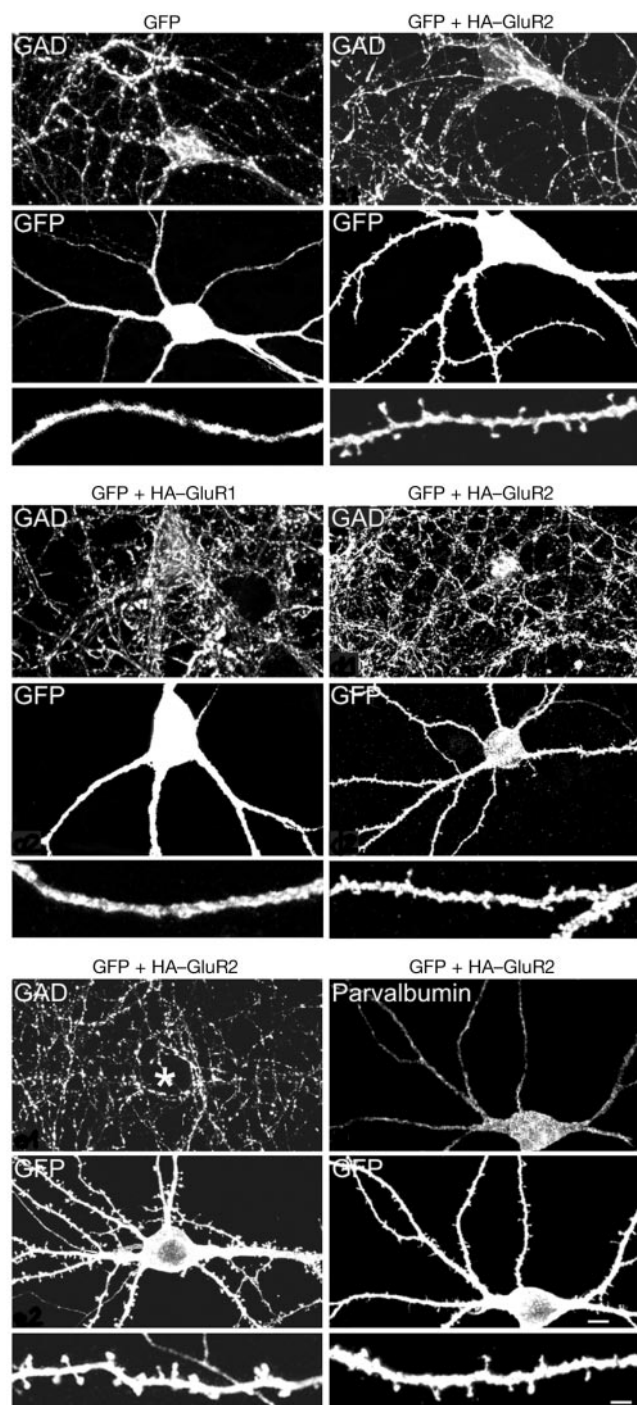


Figure 4 GluR2 induces dendritic spines in GABA-releasing interneurons. Hippocampal neurons at DIV14 were transfected with EGFP alone, or with EGFP plus the indicated HA-GluR constructs, and stained for GFP, GAD and HA at DIV22. Paired panels show GFP and GAD staining in the same neuron, as indicated. Bottom panels show higher magnification of the dendrite (GFP image). Interneurons (defined by typical dendritic morphology and GAD6 immunoreactivity in cell soma) have no dendritic spines when transfected with EGFP (top panel, left images) or HA-GluR1 (middle panel, left images), but display protrusions with a clear head and neck when transfected with HA-GluR2 (top and middle panels, right images). The left images of the bottom panel are an example of a presumptive pyramidal neuron (GAD-negative cell soma) transfected with GFP plus HA-GluR2. The right images of the bottom panel are an example of an interneuron (defined by parvalbumin immunoreactivity) transfected with EGFP plus HA-GluR2. Scale bar, 10 μ m (low magnification) and 2.5 μ m (high magnification).

or undetectable levels of endogenous GluR2 (refs 11,12). Thus, it is possible that GluR2 is a determinant for dendritic spine formation.

A surprising finding is that the spine-promoting determinant lies within the extracellular NTD of GluR2. The activity of the GluR2-NTD seems to be mediated through specific interactions that can be 'competed' by soluble Fc fusion protein of the NTD. The NTD of GluR2, similar to that of other glutamate receptors, is distantly related to bacterial periplasmic binding proteins (PBP), in particular LIVBP (leucine isoleucine valine binding protein)^{9,13,14}. A second more C-terminal PBP-like domain forms the glutamate-binding domain in AMPA receptors^{15,16}.

Little is known about the structure or physiological function of the NTD of AMPA receptors. In NMDA receptor subunits, the NTD has been implicated in modulation of NMDA receptor activity: for example, glycine-independent desensitization^{17,18}, proton sensitivity¹⁴ and zinc binding/inhibition^{13,19,20}. However, few extracellular modulators have been found for AMPA receptors. The NTD seems to be unimportant for surface expression or basic channel functions of ionotropic glutamate receptors^{20–22}. So far, the only function ascribed to the NTD of AMPA receptors is in dimerization and subtype-specific assembly of subunits^{21,23}. However, our findings cannot be easily explained by the requirement of the NTD in subunit assembly. For instance, both NTDR2/R1 and NTDR1/R2 chimaeras should be competent for subunit assembly, but only NTDR2/R1 is active in spine morphogenesis.

Our study points to a new function of the NTD of AMPA receptor subunits. We propose that the NTD of GluR2 is the specific site for an extracellular protein–protein interaction that is important for the formation, growth and/or maintenance of dendritic spines. The NTD of GluR2 could be a receptor for a secreted factor that signals directly through GluR2 to promote spine morphogenesis, perhaps through a 'metabotropic' function of AMPA receptors²⁴. Alternatively the GluR2-NTD might bind to a specific membrane protein in *cis*—analogous to the interaction of NMDA receptors and EphB2 receptor tyrosine kinase²⁵. In such a model, GluR2, by means of its NTD, recruits to the postsynaptic membrane another transmembrane protein that is more directly involved in spine morphogenesis. A third hypothesis is that GluR2-NTD is involved in a trans-synaptic interaction by binding directly to the extracellular domain of a presynaptic membrane protein (analogous to the neuroligin–neurexin interaction involved in synapse development²⁶). We note that if the presynaptic binding partner of GluR2-NTD is expressed specifically on glutamate-releasing neurons, such a trans-synaptic protein interaction would offer a simple way to align glutamate receptors with glutamate-releasing terminals. □

Methods

DNA constructs

HA epitope tag was inserted three amino acids C-terminal of the signal peptide of GluR subunits (before Gly-31 in GluR2). The deletion constructs were made by polymerase chain reaction (PCR) amplification with appropriate oligonucleotides. Chimaeras NTDR1/R2 and NTDR2/R1 were made by fusing the NTD (amino acids 1–400) of GluR1 to amino acid 402 of GluR2, and NTD (1–400) of GluR2 to amino acid 401 of GluR1. GluR1/R2C and GluR2/R1C were made by fusing the cytoplasmic tail of GluR2 (residues 820–862) to amino acid 812 of GluR1, and the tail of GluR1 (812–889) to amino acid 820 of GluR2. Point mutations were made with the QuikChange system (Stratagene). HA-GluR constructs were expressed in mammalian expression vector GW1 (CMV promoter).

Cell cultures and transfection

Primary hippocampal cultures were prepared from embryonic day 18–19 rat brains²⁷ and grown in neurobasal medium supplemented with B27. Neurons were plated on coverslips coated with poly-D-lysine (30 μ g ml^{–1}) and laminin (2 μ g ml^{–1}) at a density of 75,000 per well, and transfected by the calcium phosphate method.

Immunostaining and antibodies

Six to eight days after transfection, neurons were fixed and immunostained as described²⁸. Staining of endogenous surface AMPA receptors has been described previously²⁹.

Antibodies/probes used were: rabbit anti-HA (Santa Cruz Biotechnology); Shank guinea-pig antibody 1123 (gift from E. Kim); bassoon mouse monoclonal (Stressgen); anti-GAD monoclonal (GAD6; Developmental Studies Hybridoma Bank); anti-GluR2 monoclonal (Chemicon International); anti-parvalbumin monoclonal (Sigma); Texas-red-conjugated phalloidin and Alexa-488 and Alexa-568 secondary antibodies (Molecular Probes).

Fc-NTD fusion protein inhibition

Amino acids 1–404 of GluR1 (NTDR1) and 1–405 of GluR2 (NTDR2) were amplified by PCR and subcloned in-frame, N-terminal of the Fc segment of modified pJFE vector²⁵. Fc fusion proteins were expressed in HEK293T cells, purified from culture medium by protein A Sepharose Fast Flow chromatography (Pharmacia), eluted with 100 mM glycine (pH 2.5), neutralized by adding Tris-HCl pH 8.5, and dialysed against 20 mM HEPES pH 7.4, 150 mM NaCl. Purified proteins were >90% pure as assessed by SDS–polyacrylamide gel electrophoresis. For the Fc–NTD transfection experiment, neurons (DIV14) were co-transfected with EGFP and Fc–NTD constructs. At DIV22, neurons were fixed and stained with anti-human Fc Alexa-568 antibody. The Fc–NTD constructs were expressed using vector β -actin –16-pl (β -actin promoter).

GluR2 siRNA

For plasmid-based RNA inhibition of GluR2, the following complementary oligonucleotides were annealed and inserted into the *HindIII/BglII* sites of pSUPER vector (Oligo Engine³⁰): 5'-GATCCCGGAGCACTCTTAGCTTGATTCAGAGATCAAGCTAAGGAGTGCTCC TTTTGGAAA-3', and 5'-AGCTTTTCCAAAAGGAGCACTCC TAGCTTGATCTCTTGAATCAAGCTAAGGAGTGCTCCGGG-3' (corresponding to nucleotides 400–418 of rat GluR2). The specificity and efficacy of this construct to interfere with GluR2 expression was first tested against heterologously expressed GluR in COS-7 cells.

Image acquisition and quantification

Labelled transfected neurons were chosen randomly for quantification from four coverslips from five independent experiments for each construct. Fluorescence images were acquired with a Biorad MRC1024 confocal microscope, using a Nikon $\times 60$ objective with sequential acquisition setting at 1,280 $\times$ 1,024 pixel resolution. All morphometric measurements were made with Metamorph image analysis software (Universal Imaging Corporation). Fluorescence intensity of staining of endogenous synaptic proteins was measured as described²⁸. Both image acquisition and morphometric quantification were performed by investigators who were 'blind' to the experimental condition. All measurements are given as mean $\pm$ standard error of the mean (s.e.m.).

Received 31 December 2002; accepted 28 April 2003; doi:10.1038/nature01781.

- Matsuzaki, M. *et al.* Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. *Nature Neurosci.* **4**, 1086–1092 (2001).
- Nusser, Z. AMPA and NMDA receptors: similarities and differences in their synaptic distribution. *Curr. Opin. Neurobiol.* **10**, 337–341 (2000).
- Nusser, Z. *et al.* Cell type and pathway dependence of synaptic AMPA receptor number and variability in the hippocampus. *Neuron* **21**, 545–559 (1998).
- Takumi, Y., Ramirez-Leon, V., Laake, P., Rinovik, E. & Ottersen, O. P. Different modes of expression of AMPA and NMDA receptors in hippocampal synapses. *Nature Neurosci.* **2**, 618–624 (1999).
- Engert, F. & Bonhoeffer, T. Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399**, 66–70 (1999).
- Maletic-Savatic, M., Malinow, R. & Svoboda, K. Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* **283**, 1923–1927 (1999).
- Shi, S. H. *et al.* Rapid spine delivery and redistribution of AMPA receptors after synaptic NMDA receptor activation. *Science* **284**, 1811–1816 (1999).
- Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R. & Muller, D. LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402**, 421–425 (1999).
- O'Hara, P. J. *et al.* The ligand-binding domain in metabotropic glutamate receptors is related to bacterial periplasmic binding proteins. *Neuron* **11**, 41–52 (1993).
- Shi, S., Hayashi, Y., Esteban, J. A. & Malinow, R. Subunit-specific rules governing AMPA receptor trafficking to synapses in hippocampal pyramidal neurons. *Cell* **105**, 331–343 (2001).
- Jonas, P., Racca, C., Sakmann, B., Seeburg, P. H. & Monyer, H. Differences in Ca^{2+} permeability of AMPA-type glutamate receptor channels in neocortical neurons caused by differential GluR-B subunit expression. *Neuron* **12**, 1281–1289 (1994).
- Washburn, M. S., Numberger, M., Zhang, S. & Dingleline, R. Differential dependence on GluR2 expression of three characteristic features of AMPA receptors. *J. Neurosci.* **17**, 9393–9406 (1997).
- Paoletti, P. *et al.* Molecular organization of a zinc binding N-terminal modulatory domain in a NMDA receptor subunit. *Neuron* **28**, 911–925 (2000).
- Masuko, T. *et al.* A regulatory domain (R1–R2) in the amino terminus of the N-methyl-D-aspartate receptor: effects of spermine, protons, and ifenprodil, and structural similarity to bacterial leucine/isoleucine/valine binding protein. *Mol. Pharmacol.* **55**, 957–969 (1999).
- Stern-Bach, Y. *et al.* Agonist selectivity of glutamate receptors is specified by two domains structurally related to bacterial amino acid-binding proteins. *Neuron* **13**, 1345–1357 (1994).
- Armstrong, N., Sun, Y., Chen, G.-Q. & Gouaux, E. Structure of a glutamate-receptor ligand-binding core in complex with kainate. *Nature* **395**, 913–917 (1998).
- Krupp, J. J., Vissel, B., Heinemann, S. F. & Westbrook, G. L. N-terminal domains in the NR2 subunit control desensitization of NMDA receptors. *Neuron* **20**, 317–327 (1998).
- Villarroel, A., Regalado, M. P. & Lerma, J. Glycine-independent NMDA receptor desensitization: localization of structural determinants. *Neuron* **20**, 329–339 (1998).
- Choi, Y. B. & Lipton, S. A. Identification and mechanism of action of two histidine residues underlying high-affinity Zn^{2+} inhibition of the NMDA receptor. *Neuron* **23**, 171–180 (1999).

- Fayyazuddin, A., Villarreal, A., Le Goff, A., Lerma, J. & Neyton, J. Four residues of the extracellular N-terminal domain of the NR2A subunit control high-affinity Zn^{2+} binding to NMDA receptors. *Neuron* **25**, 683–694 (2000).
- Ayalon, G. & Stern-Bach, Y. Functional assembly of AMPA and kainate receptors is mediated by several discrete protein-protein interactions. *Neuron* **31**, 103–113 (2001).
- Pasternack, A. *et al.* Alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor channels lacking the N-terminal domain. *J. Biol. Chem.* **277**, 49662–49667 (2002).
- Leuschner, W. D. & Hoch, W. Subtype-specific assembly of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptor subunits is mediated by their N-terminal domains. *J. Biol. Chem.* **274**, 16907–16916 (1999).
- Hayashi, T., Umemori, H., Mishina, M. & Yamamoto, T. The AMPA receptor interacts with and signals through the protein tyrosine kinase Lyn. *Nature* **397**, 72–76 (1999).
- Dalva, M. B. *et al.* EphB receptors interact with NMDA receptors and regulate excitatory synapse formation. *Cell* **103**, 945–956 (2000).
- Scheiffele, P., Fan, J., Choih, J., Fetter, R. & Serafini, T. Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell* **101**, 657–669 (2000).
- Brewer, G. J., Torricelli, J. R., Evege, E. K. & Price, P. J. Optimized survival of hippocampal neurons in B27-supplemented Neurobasal, a new serum-free medium combination. *J. Neurosci. Res.* **35**, 567–576 (1993).
- Sala, C. *et al.* Regulation of dendritic spine morphology and synaptic function by Shank and Homer. *Neuron* **31**, 115–130 (2001).
- Passafium, M., Piech, V. & Sheng, M. Subunit-specific temporal and spatial patterns of AMPA receptor exocytosis in hippocampal neurons. *Nature Neurosci.* **4**, 917–926 (2001).
- Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to F. Beretta and P. Grossano for technical support. The work was supported by Telethon-Italy (M.P.), Human Frontier Science Program long-term fellowship (T.N.), the Giovanni Armenise-Harvard Foundation Career Development Program (C.S.) and NIH (M.S.). M.S. is an Investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.S. (msheng@mit.edu) or M.P. (m.passafium@in.cnr.it).

Accelerated vaccination for Ebola virus haemorrhagic fever in non-human primates

Nancy J. Sullivan¹, Thomas W. Geisbert², Joan B. Geisbert², Ling Xu¹, Zhi-yong Yang¹, Mario Roederer¹, Richard A. Koup¹, Peter B. Jahrling² & Gary J. Nabel¹

¹Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bldg. 40, Room 4502, MSC 3005, 40 Convent Drive, Bethesda, Maryland 20892-3005, USA

²United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland 21702-5011, USA

Containment of highly lethal Ebola virus outbreaks poses a serious public health challenge. Although an experimental vaccine has successfully protected non-human primates against disease¹, more than six months was required to complete the immunizations, making it impractical to limit an acute epidemic. Here, we report the development of accelerated vaccination against Ebola virus in non-human primates. The antibody response to immunization with an adenoviral (ADV) vector encoding the Ebola glycoprotein (GP) was induced more rapidly than with DNA priming and ADV boosting, but it was of lower magnitude. To determine whether this earlier immune response could nonetheless protect against disease, cynomolgus macaques were challenged with Ebola virus after vaccination with ADV–GP and nucleoprotein (NP) vectors. Protection was highly effective and correlated with the generation of Ebola-specific CD8⁺ T-cell

and antibody responses. Even when animals were immunized once with ADV-GP/NP and challenged 28 days later, they remained resistant to challenge with either low or high doses of virus. This accelerated vaccine provides an intervention that may help to limit the epidemic spread of Ebola, and is applicable to other viruses.

Mice were immunized with plasmid DNA encoding Ebola GP, the trimeric virion-associated glycoprotein² involved in cellular pathogenicity^{3–6}, followed by boosting with ADV-GP, or with ADV-GP only. The antibody response, a surrogate for protection^{1,7}, was measured using an enzyme-linked immunosorbent assay (ELISA). After DNA vaccination, titres were modest but increased 100- to 1,000-fold with ADV-GP boosting (Fig. 1a). In contrast, vaccination with ADV-GP gave rise to a lower antibody titre, but it was generated more rapidly. To investigate whether immunization with adenoviral vectors alone might protect against Ebola

virus infection, alternative immunization schedules in macaques were developed for comparison to the previous DNA/ADV protocol (Fig. 1b, middle and bottom panels compared with top panel).

Cynomolgus macaques were immunized with ADV-GP and ADV-NP, followed by boosting 9 weeks later (Fig. 1b, middle panel). One week after the boost, animals were challenged with either a low (13 plaque-forming units (PFUs)) or high (1,500 PFUs) dose of a 1995 isolate of Ebola virus Zaire. These doses were uniformly fatal 6–12 days afterwards in saline-injected control animals. In contrast, the ADV-GP/NP immunized monkeys ($n = 4$) were completely protected, confirmed by viral load (Fig. 2). Analysis of the cell-mediated and humoral immune responses revealed significant increases in the CD8⁺ T-cell response to Ebola antigens by intracellular cytokine staining for interferon (IFN)- γ , seen before exposure to virus, in contrast to control animals where no response was seen (Fig. 3a). Similarly, antibody titres to the virus were stimulated in vaccinated animals, which minimally increased after the viral challenge (Fig. 3b). No substantial increases were observed in the numbers of Ebola-specific CD4⁺ T cells at this time (data not shown). Both CD8⁺ cellular and humoral immune responses therefore were associated with protection.

A second adenoviral immunization did not substantially increase the Ebola-specific immune responses (data not shown), raising the notion that the primary immunization was sufficient to confer protection. To address this possibility, a single immunization was given, and animals were challenged one month afterwards (Fig. 1b, bottom panel). Both at low and high viral challenge doses, animals were completely protected against infection (Fig. 4a). In this case, changes in the intracellular IFN- γ response in T lymphocytes were not consistently seen (data not shown); however, Ebola-specific T-cell responses were detected with intracellular tumour-necrosis factor (TNF)- α . CD8 responses were observed before challenge or were induced soon thereafter in five of eight animals, once again correlating with protection against infection (Fig. 4b, right). In contrast, CD4⁺ responses, not detectable before inoculation, increased after challenge (Fig. 4b, left). Immunoglobulin- γ (IgG) antibody titres, readily detected at the time of inoculation, were also associated with protection (Fig. 4c). These data demonstrated that a single ADV-GP/NP injection can accelerate vaccine protec-

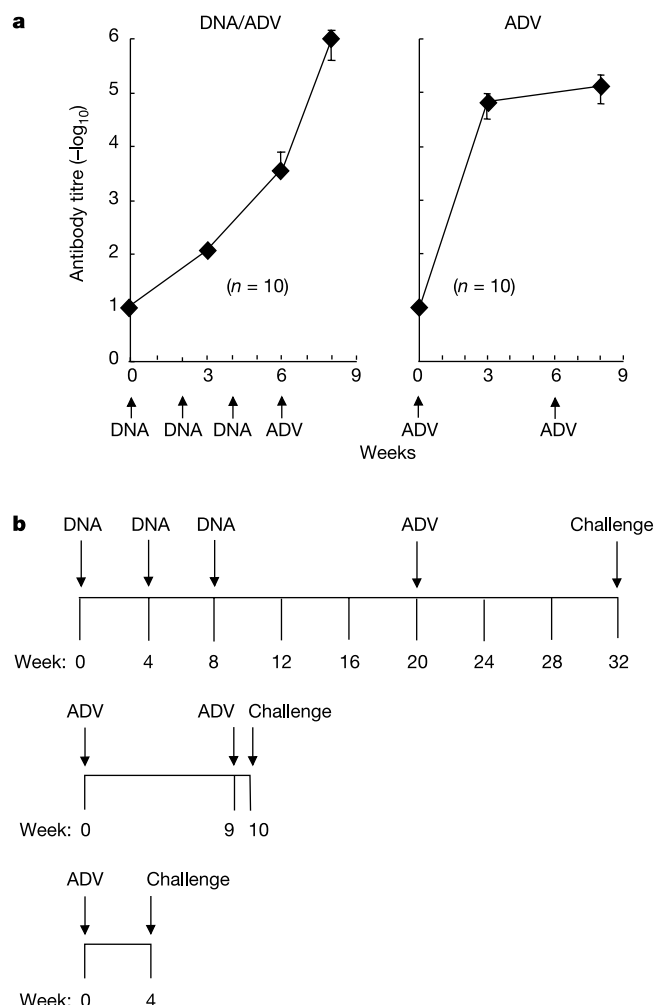


Figure 1 Comparison of the Ebola-specific antibody responses by heterologous DNA/ADV prime-boost or ADV prime-boost vaccination in mice. **a**, The time course of Ebola-specific antibody responses by DNA prime and adenovirus boost compared with adenoviral immunization alone is shown (see Methods). Data represent the relative ELISA titre to Ebola GP after immunization with DNA/ADV-GP or ADV-GP/ADV-GP in BALB/c mice using a log scale. **b**, Immunization schedule for previously used heterologous prime-boost vaccine (top), adenoviral prime and boost (middle), and single adenoviral virus (bottom) immunizations. Challenge was performed with a 1995 isolate of Ebola virus (Zaire) at 32, 10 or 4 weeks after the initial immunization, respectively.

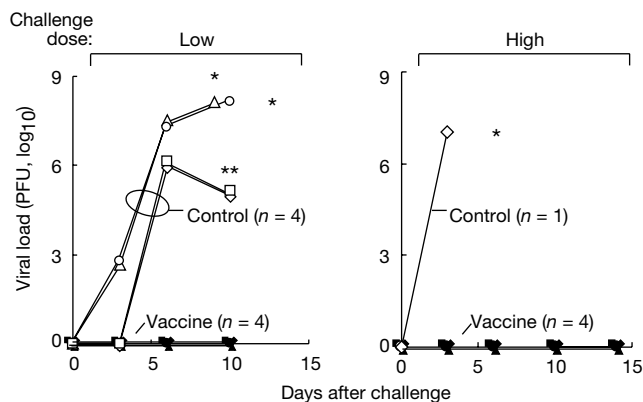


Figure 2 Protection against lethal challenge in non-human primates using adenoviral priming and boosting. Plasma viraemia in monkeys after infection with Ebola virus. Asterisks represent the time of death in control animals. The data represent the reciprocal endpoint dilution of serum for each monkey. Results are shown for four immunized animals challenged with Ebola Zaire at 13 PFUs (low dose; filled symbols, left), four immunized animals challenged at 1,500 PFUs (high dose; filled symbols, right), and five saline-injected control animals (open symbols).

tion and long-term survival against Ebola in non-human primates (Fig. 4d).

Ebola virus infection is characterized by its rapid onset, high person-to-person transmissibility, and significant mortality rate. The mainstay of treatment has been supportive therapy, and prevention has been dependent on containment using barrier precautions. Effective protection was achieved previously in primates with a heterologous DNA prime and adenoviral boost strategy. The prime-boost immunization relies on the ability of the adenoviral boost to expand the primary T-cell response induced by DNA vaccination. When animals are primed with ADV vectors alone, a robust Ebola-virus-specific cellular and humoral immune response is more rapidly achieved, although the response to a second ADV-GP/NP injection is blunted, probably because of anti-vector immunity. Here, we explored the possibility that this more rapid initial immune response may nonetheless confer protection and outweigh the stronger immune response that requires additional time. A single immunization with an adenoviral vector encoding Ebola virus proteins is sufficient to confer protection against lethal challenge within four weeks, and this response correlates with both cellular and humoral immune responses to the infection.

Although antibody titres correlated here with the protective response, previous studies in non-human primates have suggested that the passive transfer of antibody is insufficient to provide long-lasting protection against Ebola virus⁸. In rodent studies with adapted Ebola virus, passive transfer of antibodies^{9,10} or adoptive transfer of cytotoxic T cells¹¹ showed protection when given before infection. A more sensitive but less quantitative CD4 lymphoproliferative response correlated with protection in the previous DNA/ADV prime-boost study, in which CD8 responses were not measured¹. In addition to the antibody response induced by the

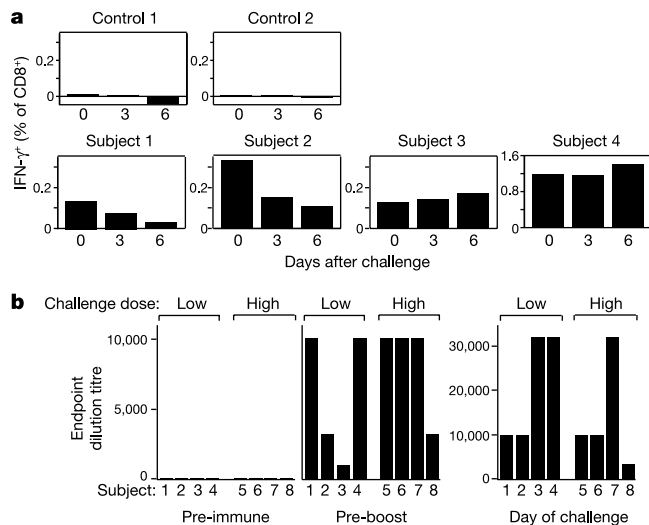


Figure 3 Immune responses to adenoviral prime and boost vaccination in cynomolgus macaques. **a**, Intracellular flow cytometry was performed to quantify IFN- γ production from Ebola-specific CD8 lymphocytes from saline injected (control) or ADV-GP/NP immunized (subject) monkeys at weeks 0 and 9. Immune responses before (day 0) and after (days 3, 6) challenge at week 10 are shown for CD8 cells. No substantial increases were observed in the CD4 population. Non-stimulated cells gave responses similar to those of the control subjects, at background levels. The gating strategy for FACS analysis is similar to previous analyses (Supplementary Information). **b**, ELISA titres of Ebola-specific antibodies in serum of vaccinated animals collected at week 0 (pre-immune, left), week 9 (pre-boost, middle) and week 10 (day of challenge, right) relative to the time of the first immunization. ELISA results represent endpoint dilution titres determined by optical density as described in Methods.

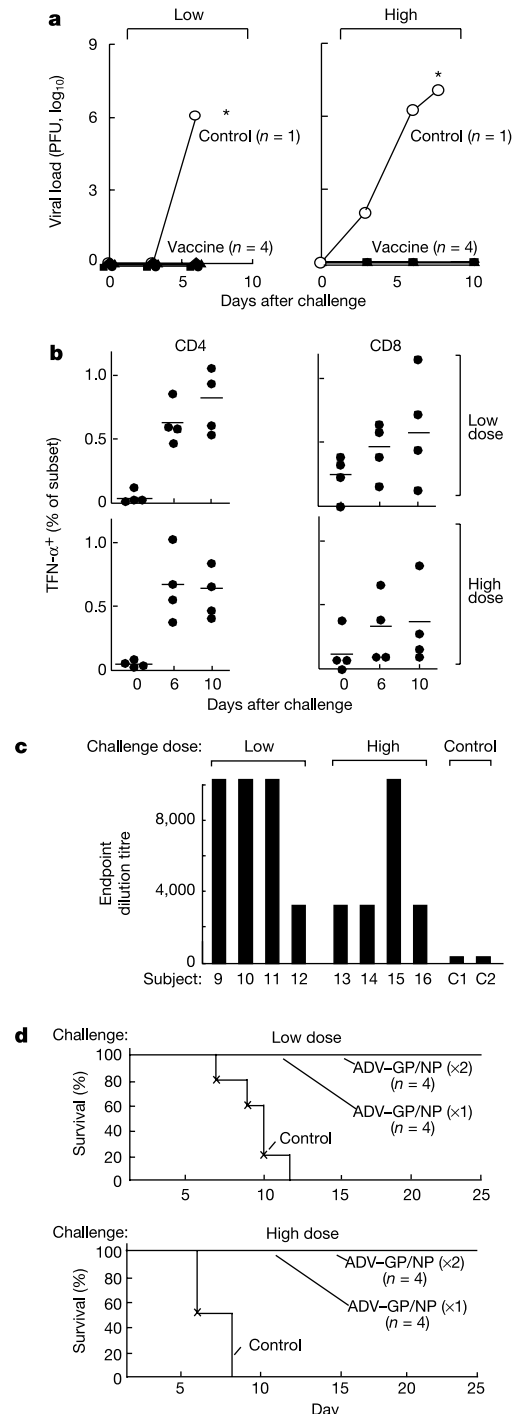


Figure 4 Protection against lethal challenge in non-human primates using a single adenoviral immunization. **a**, Immunization and challenge were performed with the 1995 Zaire subtype Ebola virus as in Fig. 1b (bottom), and plasma viraemia in monkeys after challenge was measured as above (see Fig. 2 legend) for four immunized animals inoculated with 18 PFUs (low dose; filled symbols, left) and four animals injected with 1,762 PFUs (high dose; filled symbols, right) or two saline-injected controls (open symbols). **b**, Intracellular flow cytometry was performed using antibodies to TNF- α in CD4 and CD8 lymphocytes from immunized monkeys (subject), each panel representing an individual macaque. Immune responses before (day 0) and after (days 6, 10) challenge on day 28 are shown. Horizontal bars indicate the average value per group, and filled circles represent values for individual subjects. **c**, Endpoint dilution ELISA titres of Ebola-specific antibodies in serum collected two weeks after immunization with ADV-GP/NP, determined by optical density as described in Methods. **d**, Kaplan-Meier survival curve of macaques, immunized as indicated, and challenged with a low or high dose of PFUs of Ebola virus.

vaccine in the present study, both CD4 and CD8 responses were observed after the challenge. The fact that CD4 responses were not observed before challenge in either protocol whereas CD8 responses were more consistently seen beforehand suggests that the CD8 response is likely to have an important role in protection in non-human primates, but further analysis will be required to assess the relative importance of the cellular and humoral immune responses in the mechanism of protection.

The approach to single vaccine injection with ADV vectors is relevant to the containment and treatment of Ebola virus and related outbreaks that are continuing to emerge in central Africa¹². Should this vaccine approach prove effective in humans, our finding raises the possibility that ring vaccination could be used to contain outbreaks, similar to smallpox in the past. This result also suggests alternative strategies for vaccination against Ebola or other acute pathogenic diseases. The prime-boost strategy remains more immunologically potent and, if the response is highly durable, may still be useful for preventative vaccines, for example, in hospital workers. In contrast, the single adenoviral vaccine administration may be better used during acute outbreaks. It is also possible that alternative viral vectors, such as those derived from other adenovirus serotypes or from poxvirus vectors, might be used to boost an ADV type 5 vector primary immunization. Alternative ADV serotypes will also help to overcome immunity to natural ADV type 5 infection that could potentially reduce vaccine efficacy in some populations. A one-shot vaccine may be helpful in the control of Ebola virus outbreaks in great ape populations of central Africa¹³. Analogous single-dose ADV vector immunization could also be used for other emerging and highly lethal infectious pathogens, such as Marburg, Lassa or the SARS coronavirus. □

Methods

Vector construction

ADV-GP and ADV-NP were prepared as described previously¹. The recombinant adenoviral vector was made according to previously published methods¹⁴. A dose of 10^{10} (mice) or 10^{12} (non-human primates) adenoviral vector particles for each component was administered to each animal without adverse effects.

Animal study and safety

Twenty cynomolgus macaques (*Macaca fascicularis*), 3 yr old and weighing 2–3 kg, obtained from Covance, were used for immunization and challenge experiments. The monkeys, housed singly, were anaesthetized with ketamine to obtain blood specimens and to administer vaccines. They received regular enrichment according to the Guide for the Care and Use of Laboratory Animals (DHEW number NIH 86-23). Before Ebola virus challenge and to the end of each experiment, the animals were maintained in the Maximum Containment Laboratory (BSL-4) and fed and checked daily.

Mouse immunization

DNA and adenoviral vectors expressing Ebola Zaire glycoprotein (Mayinga strain) were constructed as described previously^{7,15} with gene expression under control of the cytomegalovirus enhancer and promoter in the plasmid. Mice ($n = 10$ per group) were immunized intramuscularly with 100 µg DNA (pGP) and/or 10^{10} particles of adenovirus (ADV-GP). DNA vaccination was performed on days 0, 14 and 24 with adenoviral boost on day 42. Adenoviral injection was performed on days 0 and 42, and samples were collected for ELISA titres at the indicated times. ELISA IgG titres were determined using 96-well plates as previously described¹⁶, and specific antigen binding was detected using a goat anti-human IgG (H + L)-horseradish conjugate and ABTS/peroxide (substrate/indicator).

ELISA

Polyvinyl chloride ELISA plates (Dynatech) were coated with 50 µl antigen per well and incubated overnight at 4°C. All further incubations were carried out at room temperature. The antigen used was purified Ebola virus (about 1 mg ml⁻¹ total protein) inactivated by gamma irradiation. Plates were then washed five times with PBS containing Tween-20. Test sera were diluted in half-log concentrations from 1:31.6 through to 1:100,000 and allowed to react with the antigen-coated wells for 60 min. After washing plates five times, goat anti-monkey IgG (whole molecule; ICN Biomedicals) conjugated to horseradish peroxidase was used as a detection antibody. Bound IgG was detected by 2,2'-azino-bis-[3-ethylbenzothiazoline-6-sulphonic acid] diammonium salt and the optical density was determined. A panel of normal serum was run each time the assay was performed. A cut-off value for a positive result was calculated as the mean optical density (at a 1:100 dilution) for the normal sera plus 3 standard deviations.

Intracellular cytokine analysis

Peripheral blood mononuclear cells were isolated from cynomolgus macaque whole-blood samples by separation over Ficoll. Approximately 1×10^6 cells were stimulated in 200 µl RPMI medium (GIBCO) for 6 h at 37°C with anti-CD28 and anti-CD49d antibodies and either DMSO or a pool of 15-nucleotide peptides spanning the Ebola GP Zaire (Mayinga strain) open reading frame. The peptides were 15 nucleotides overlapping by 11 spanning the entire Ebola glycoprotein at a final concentration of 2 µg ml⁻¹. Cells were fixed and permeabilized with FACS lyse (Becton Dickinson) supplemented with Tween-20, and stained with a mixture of antibodies against lineage markers (CD3-PE, CD4-PerCP, CD8-FITC) and either TNF-APC or IFN-γ-APC. Samples were run on a FACS Calibur and analysed using the software FlowJo. Positive gating for lymphocytes using forward versus side scatter was followed by CD3⁺/CD8⁺ and CD3⁺/CD4⁺ gating, and specific populations were further defined by anti-CD4 and anti-CD8 positivity, respectively. Cytokine-positive cells were defined as a percentage within these individual lymphocyte subsets, and at least 200,000 events were analysed for each sample.

Macaque immunization

In conducting this research, the investigators adhered to the Guide for the Care and Use of Laboratory Animals, prepared by the Institute of Laboratory Animal Resources, National Research Council. The facilities are fully accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International. Cynomolgus macaques were injected intramuscularly at the indicated times (Fig. 1b) with an equal mixture of 2×10^{12} particles of ADV-GP and ADV-NP. Viral challenge was performed by inoculation of animals in the left or right caudal thigh with 0.5 ml of viral stock that contained a target dose of either about 10 or about 1,000 PFUs of Ebola virus (Zaire species) at ten weeks (Fig. 2) or four weeks (Fig. 4) after the initial immunization, and actual titre was confirmed by plaquing. No adverse effects of the adenovirus vaccination were observed acutely. The Ebola virus used in this study was originally obtained from a fatally infected human from the former Zaire in 1995 (ref. 17). Collection of serum and blood for viral load and ELISA titres was performed as previously described¹.

Received 22 May; accepted 30 June 2003; doi:10.1038/nature01876.

- Sullivan, N. J., Sanchez, A., Rollin, P. E., Yang, Z.-Y. & Nabel, G. J. Development of a preventive vaccine for Ebola virus infection in primates. *Nature* **408**, 605–609 (2000).
- Sanchez, A. *et al.* Biochemical analysis of the secreted and virion glycoproteins of Ebola virus. *J. Virol.* **72**, 6442–6447 (1998).
- Yang, Z.-Y. *et al.* Identification of the Ebola virus glycoprotein as the main viral determinant of vascular cell cytotoxicity and injury. *Nature Med.* **6**, 886–889 (2000).
- Volchkov, V. E. *et al.* Recovery of infectious Ebola virus from complementary DNA: RNA editing of the GP gene and viral cytotoxicity. *Science* **291**, 1965–1969 (2001).
- Chan, S. Y., Ma, M. C. & Goldsmith, M. A. Differential induction of cellular detachment by envelope glycoproteins of Marburg and Ebola (Zaire) viruses. *J. Gen. Virol.* **81**, 2155–2159 (2000).
- Takada, A. *et al.* Downregulation of β1 integrins by Ebola virus glycoprotein: implication for virus entry. *Virology* **278**, 20–26 (2000).
- Xu, L. *et al.* Immunization for Ebola virus infection. *Nature Med.* **4**, 37–42 (1998).
- Jahrling, P. B. *et al.* Evaluation of immune globulin and recombinant interferon-α2b for treatment of experimental Ebola virus infections. *J. Infect. Dis.* **179**, S224–S234 (1999).
- Parren, P. W., Geisbert, T. W., Maruyama, T., Jahrling, P. B. & Burton, D. R. Pre- and postexposure prophylaxis of Ebola virus infection in an animal model by passive transfer of a neutralizing human antibody. *J. Virol.* **76**, 6408–6412 (2002).
- Gupta, M., Mahanty, S., Bray, M., Ahmed, R. & Rollin, P. E. Passive transfer of antibodies protects immunocompetent and immunodeficient mice against lethal Ebola virus infection without complete inhibition of viral replication. *J. Virol.* **75**, 4649–4654 (2001).
- Wilson, J. A. & Hart, M. K. Protection from Ebola virus mediated by cytotoxic T lymphocytes specific for the viral nucleoprotein. *J. Virol.* **75**, 2660–2664 (2001).
- World Health Organization. Disease Outbreaks. Ebola Haemorrhagic Fever (<http://www.who.int/disease-outbreak-news/disease/A98.4.htm>) (2003).
- Walsh, P. D. *et al.* Catastrophic ape decline in western equatorial Africa. *Nature* **422**, 611–614 (2003).
- Aoki, K., Barker, C., Danthinne, X., Imperiale, M. J. & Nabel, G. J. Efficient generation of recombinant adenoviral vectors by Cre-lox recombination in vitro. *Mol. Med.* **5**, 224–231 (1999).
- Ohno, T. *et al.* Gene therapy for vascular smooth muscle cell proliferation after arterial injury. *Science* **265**, 781–784 (1994).
- Ksiazek, T. G. *et al.* Enzyme immunoassay for Ebola virus antigens in tissues of infected primates. *J. Clin. Microbiol.* **30**, 947–950 (1992).
- Jahrling, P. B. *et al.* Passive immunization of Ebola virus-infected cynomolgus monkeys with immunoglobulin from hyperimmune horses. *Arch. Virol. Suppl.* **11**, 135–140 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Suhana and A. Tislerics for help with manuscript preparation; K. Stroud and T. Miller for figure preparation; D. Levy for helping to define conditions for intracellular cytokine analysis; L. Hensley of USAMRIID and members of the Nabel laboratory for discussions; and T. Larsen for pathology support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.J.N. (gnabel@nih.gov).

Exclusion of germ plasm proteins from somatic lineages by cullin-dependent degradation

Cynthia DeRenzo^{1*}, Kimberly J. Reese^{1*†} & Geraldine Seydoux¹

¹Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

* These authors contributed equally to this work

† Present address: HHMI, Department of Cell and Developmental Biology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

In many animals, establishment of the germ line depends on segregation of a specialized cytoplasm, or 'germ plasm', to a small number of germline precursor cells during early embryogenesis¹. Germ plasm asymmetry involves targeting of RNAs and proteins to a specific region of the oocyte and/or embryo². Here we demonstrate that germ plasm asymmetry also depends on degradation of germline proteins in non-germline (somatic) cells. We show that five CCCH finger proteins, components of the *Caenorhabditis elegans* germ plasm, are targeted for degradation by the novel CCCH-finger-binding protein ZIF-1. ZIF-1 is a SOCS-box protein that interacts with the E3 ubiquitin ligase subunit elongin C. Elongin C, the cullin CUL-2, the ring finger protein RBX-1 and the E2 ubiquitin conjugation enzyme UBC5 (also known as LET-70) are all required *in vivo* for CCCH finger protein degradation. Degradation is activated in somatic cells by the redundant CCCH finger proteins MEX-5 and MEX-6, which are counteracted in the germ line by the PAR-1 kinase. We propose that segregation of the germ plasm involves both stabilization of germline proteins in the germ line and cullin-dependent degradation in the soma.

Establishment of the germ line in *C. elegans* begins with a series of unequal divisions that divide successive germline blastomeres into somatic and germline daughter cells (Fig. 1a). The germ plasm contains RNA-binding proteins and RNA-rich organelles (P granules), which are maintained preferentially in germline blastomeres. Among germ plasm components are five proteins that share two CCCH fingers: PIE-1, POS-1, MEX-1 and the functionally redundant and 70% identical MEX-5 and MEX-6 (refs 3–6). CCCH fingers (a class of zinc finger) were first described in the vertebrate protein TTP/Nup475/Tis-11, where they are implicated in RNA binding⁷. In germline blastomeres, CCCH finger proteins are on P granules and diffuse throughout the cytoplasm. During each asymmetric division, PIE-1, POS-1 and MEX-1 segregate preferentially with the germline daughter, whereas MEX-5 and MEX-6 segregate preferentially with the somatic daughter^{3–6,8,9} (Fig. 1a; see also [ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003](http://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003)). After division, the CCCH finger proteins persist in the germline blastomeres at either high (PIE-1, MEX-1 and POS-1) or low (MEX-5 and MEX-6) levels. In contrast, in somatic blastomeres they do not persist beyond one (PIE-1, MEX-1 and POS-1) or two (MEX-5 and MEX-6) divisions^{3–6,8,9} (Fig. 1a; see also movies at [ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003](http://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003)). Earlier studies with PIE-1 demonstrated that asymmetric partitioning to germline blastomeres and turnover in somatic blastomeres are mediated by two independent mechanisms¹⁰. Turnover depends on the first CCCH finger (ZF1) in PIE-1, which targets PIE-1 for degradation specifically in somatic blastomeres¹⁰. ZF1s from PIE-1, POS-1 and MEX-1 (PIE-1^{ZF1}, POS-1^{ZF1}, MEX-1^{ZF1}), and ZF2 from MEX-5 (MEX-5^{ZF2}) are sufficient to target green fluorescent protein (GFP) for degradation in somatic blastomeres (Methods; see also ref. 10 and [ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003](http://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003)). In contrast, PIE-1^{ZF2}, POS-1^{ZF2}, MEX-1^{ZF2}, MEX-5^{ZF1} and either finger in TTP/Nup475/Tis-11 are not sufficient to trigger soma-specific degradation¹⁰. These findings suggest

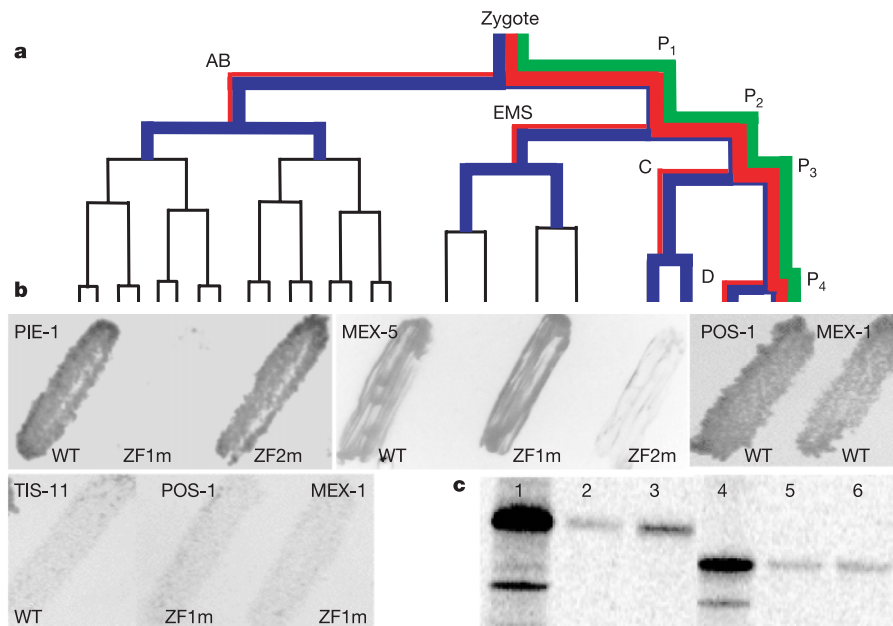


Figure 1 ZIF-1 interacts with CCCH fingers targeted for degradation. **a**, Early embryonic lineage. P₁ to P₄ are the germline blastomeres. AB, EMS, C and D are the somatic blastomeres. PAR-1 (green), PIE-1 (red) and MEX-5 (blue) segregations are indicated; line thickness indicates relative levels. **b**, Yeast transformed with ZIF-1 and finger domains from indicated proteins. ZF1m and ZF2m are SSCH instead of CCCH. Growth indicates

interaction. WT, wild type. **c**, Labelled PIE-1 proteins (full-length, lanes 1–3; ZF1 deleted, lanes 4–6) synthesized in rabbit reticulolysates and incubated with *E. coli*-synthesized GST (lanes 2 and 5) or GST–ZIF-1 (lanes 3 and 6) on Sepharose beads. Bound proteins were eluted and run on SDS–PAGE. 100% of input (lanes 1 and 4) was loaded on GST column; 50% of bound was loaded on gel.

that CCCH finger proteins are targeted for degradation by a mechanism that recognizes specific CCCH fingers.

To identify *trans*-acting factors in this process, we performed a yeast two-hybrid screen for proteins that bind to the PIE-1 fingers (Methods). The screen identified ten isolates of F59B2.6, a new protein that we named ZIF-1 (zinc finger interacting factor 1). In the two-hybrid assay, ZIF-1 interacted with the zinc finger domains of PIE-1, POS-1, MEX-1 and MEX-5, but not TIS-11 (Fig. 1b). The interaction was blocked by mutations in the finger required for degradation, but not by mutations in the other finger (Fig. 1b). Glutathione *S*-transferase (GST) pull-down assays confirmed the specificity of the ZIF-1–PIE-1 interaction (Fig. 1c). We conclude that ZIF-1 binds to CCCH finger motifs targeted for degradation in somatic lineages.

To determine the *in vivo* function of ZIF-1, *zif-1* activity was reduced in embryos by RNA-mediated interference (RNAi)¹¹. Unlike wild-type embryos, *zif-1(RNAi)* embryos maintained PIE-1, POS-1, MEX-1, MEX-5 and MEX-6 in somatic blastomeres (Fig. 2a–c and data not shown; see also ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003). *zif-1(RNAi)* embryos also failed to degrade GFP::PIE-1^{ZF1} and GFP::MEX-5^{ZF2} fusions in somatic lineages (ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003). Consistent with ZIF-1 acting by binding to CCCH fingers, *zif-1(RNAi)* did not affect the localization of a GFP::PIE-1 fusion lacking ZF1 (Fig. 2d) and did not affect the distribution of germline-enriched factors that do not contain CCCH fingers (PAR-2 and PGL-1; Fig. 2e). Furthermore, *zif-1(RNAi)* did not disrupt asymmetric partitioning in germline blastomeres: CCCH finger proteins were still preferentially segregated to the germline (PIE-1, POS-1, MEX-1) or somatic (MEX-5) daughter cell at each asymmetric division (Fig. 2a, b, g and data not shown; see also ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003). *pie-1* RNA, which in the wild type is maintained preferentially in germline blastomeres, was also not affected in *zif-1(RNAi)* embryos (Fig. 2f). Finally, *zif-1(RNAi)* did not block turnover of OMA-1, a CCCH finger protein degraded in all cells after the one-cell stage¹² (data not shown). We conclude that ZIF-1 is not generally required for protein degradation or asymmetric partitioning in germline blastomeres, but is essential for soma-specific degradation of CCCH finger proteins. *zif-1(RNAi)* embryos die during embryogenesis (Methods), suggesting that perdurance of CCCH finger proteins in somatic lineages interferes with embryonic development.

ZIF-1 does not share obvious homology with other proteins in the databases. To investigate the mechanism of action of ZIF-1, we conducted another yeast two-hybrid screen to identify potential ZIF-1 interactors (Methods). As expected, we recovered several isolates of PIE-1, POS-1, MEX-1, MEX-5 and MEX-6. In addition, we recovered two isolates of Y82E9BR.15. Sequence analysis of Y82E9BR.15 revealed that it encodes the *C. elegans* homologue of vertebrate elongin C (Supplementary Information). In human cells, elongin C complexes with the ubiquitin-like protein elongin B, the cullin Cul2 and the RING finger protein Rbx1 to form a multi-subunit E3 ubiquitin ligase (ECS ligase¹³). When bound to the substrate-recruitment subunit von Hippel-Lindau (VHL), ECS ligase targets hypoxia inducible factor (HIF1 α) for ubiquitination and subsequent degradation by the proteasome. Elongin C binds to VHL through a conserved motif (SOCS box) located near the carboxy terminus of VHL. Mutations in this sequence in von Hippel-Lindau patients prevent binding of VHL to elongin C and lead to abnormal stabilization of HIF1 α ¹⁴. ZIF-1 contains a potential SOCS box in its C terminus (Fig. 3). Mutation of a conserved cysteine in the SOCS box blocked ZIF-1's ability to interact with elongin C in the yeast two-hybrid assay, without affecting binding to PIE-1 (Fig. 4a). This finding suggests that, similar to VHL, ZIF-1 may function as a substrate-recruitment subunit for an ECS ligase (Fig. 4b). If so, elongin C, CUL-2, RBX-1 and the associated E2 ubiquitin-conjugating enzyme UBC5 should all be required for

ZF1-dependent degradation. Complete elimination of these proteins interferes with cell divisions¹⁵ (J. Liu, S. Vasudevan and E. T. Kipreos, submitted; and C.D., unpublished work); however, partial depletion by RNAi yields multicellular embryos that can be examined for CCCH finger degradation. In all cases, we found that degradation of PIE-1::GFP and GFP::PIE-1^{ZF1} fusions was blocked in these embryos (Fig. 4c and data not shown). This

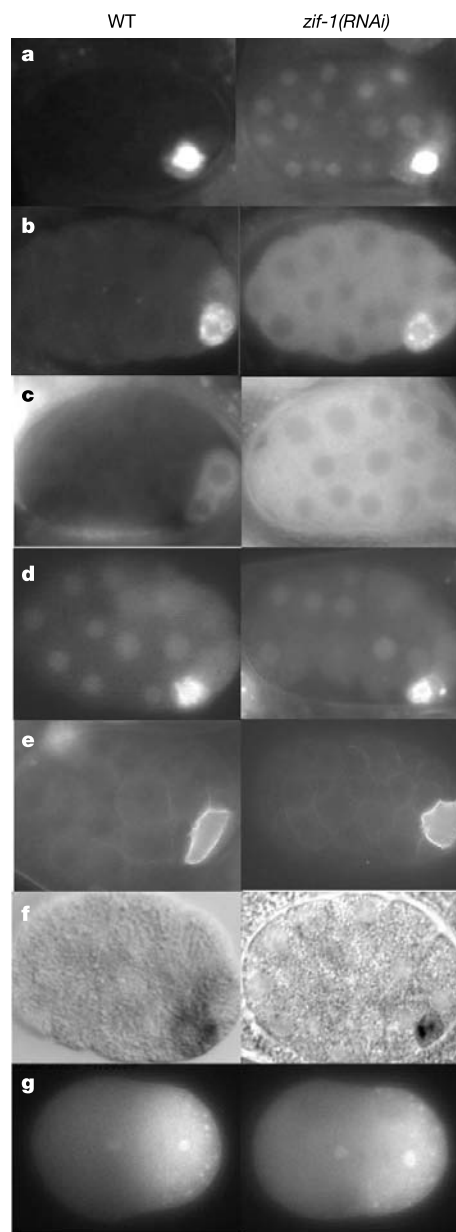


Figure 2 Depletion of ZIF-1 blocks degradation of CCCH finger proteins in somatic cells, but does not affect other soma–germline asymmetries. **a–f**, 28-cell wild-type and *zif-1(RNAi)* embryos expressing PIE-1::GFP (**a**), POS-1 (**b**, immunofluorescence⁹), GFP::MEX-5 (**c**), GFP::PIE-1 with ZF1 deleted (**d**), GFP::PAR-2 (**e**) and *pie-1* RNA (**f**, *in situ* hybridization²⁷). Measurement of GFP fluorescence (IP lab software) in ABa (soma) confirmed that PIE-1::GFP levels are higher in *zif-1(RNAi)* compared with wild type (*t*-test; $P = 0.009$), but do not change significantly in P₂ (germ line) (*t*-test; $P = 0.25$, $N = 10$ four-cell embryos for each genotype). **g**, One-cell wild-type and *zif-1(RNAi)* embryos expressing a PIE-1::GFP fusion. In later stages, the *zif-1(RNAi)* embryo failed to degrade PIE-1::GFP in somatic blastomeres (see ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003), confirming the efficient inactivation of *zif-1*.

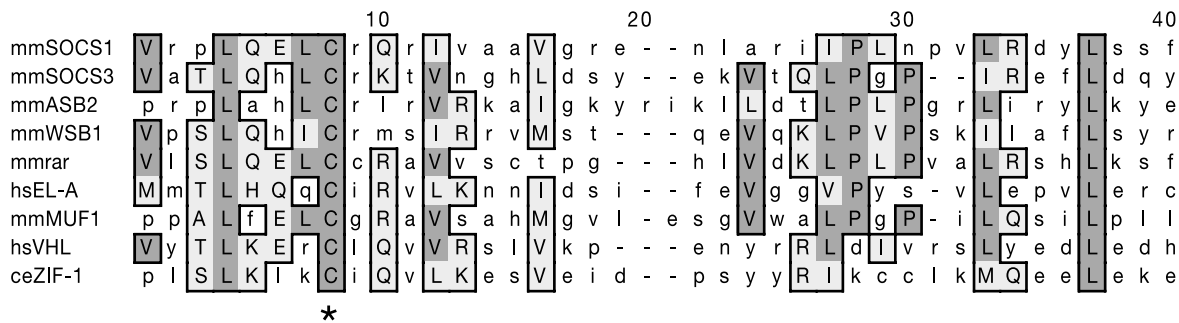


Figure 3 ZIF-1 is a SOCS-box protein. Alignment of the ZIF-1 SOCS box with SOCS boxes from other elongin C-interacting proteins¹³. Asterisk indicates cysteine mutated in Fig. 4a. ce, *C. elegans*; hs, *Homo sapiens*; mm, *Mus musculus*.

block was not due to general loss of embryonic polarity or cell viability, as the distribution of GFP::PGL-1 was unaffected (Fig. 4c). Depletion of another cullin (CUL-3) or another E2 (UBC-12) did not interfere with GFP::PIE-1^{ZF1} degradation (Supplementary Information), confirming the specific involvement of CUL-2 and UBC5. We conclude that elongin C, CUL-2, RBX-1 and UBC5 are required with ZIF-1 to degrade CCCH finger proteins in somatic cells.

F box proteins—the substrate recruitment subunits for the SCF class of ubiquitin ligases—are unstable *in vivo* and are themselves subject to ubiquitination in a SCF-dependent manner^{16–18}. Similarly, we found that a GFP::ZIF-1 fusion is unstable *in vivo*, and is stabilized when CUL-2, but not CUL-3, is inactivated (Fig. 4d; see also Supplementary Information). These observations are consistent with ZIF-1 acting as a substrate recruitment factor for a CUL-2-containing E3 ligase (Fig. 4b); however, biochemical studies will be needed to confirm this hypothesis.

ZIF-1-dependent degradation is restricted to somatic cells, as *zif-1(RNAi)* stabilizes PIE-1::GFP and GFP::PIE-1^{ZF1} levels in somatic cells without changing levels in germ cells¹⁰ (Fig. 2; see also ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003). We found that this specificity is dependent on MEX-5 and MEX-6, the redundant CCCH finger proteins that accumulate transiently in somatic blastomeres (Fig. 1a). In *mex-5(RNAi);mex-6(RNAi)*

embryos, GFP::PIE-1^{ZF1} remained stable in all blastomeres, indicating that MEX-5 and MEX-6 are required for ZIF-1-dependent degradation (Fig. 5). MEX-5 and MEX-6 levels are kept low in germline blastomeres by PAR-1, a serine threonine kinase that segregates with the germ lineage (Fig. 1a) and is essential for asymmetric divisions¹⁹. In *par-1(RNAi)* embryos, MEX-5 and MEX-6 are present at high levels in all early blastomeres⁶ and degradation of GFP::PIE-1^{ZF1} was activated in all cells (Fig. 5). This ubiquitous degradation requires MEX-5 and MEX-6, as well as ZIF-1 and CUL-2 (Fig. 5). Consistent with MEX-5 and MEX-6 triggering CCCH finger protein degradation, heat-shock-induced expression of MEX-5 in somatic blastomeres of *mex-5(zu199);mex-6(RNAi);pie-1(RNAi)* embryos correlates with loss of MEX-1 expression in those cells⁶. We conclude that high levels of MEX-5 and MEX-6 activate ZIF-1-dependent degradation in somatic blastomeres, and that PAR-1 blocks ZIF-1-dependent degradation in germline blastomeres, at least in part by keeping MEX-5 and MEX-6 levels low (Fig. 5b). Our data indicate that MEX-5 and MEX-6 are both activators and targets of ZIF-1-dependent degradation. We do not know whether MEX-5 and MEX-6 activate ZIF-1-dependent degradation directly, through their binding to ZIF-1, for example, or indirectly by regulating another factor. PAR-1, MEX-5 and MEX-6 are also required for the initial asymmetric partitioning of PIE-1, POS-1 and MEX-1 in germline blastomeres⁶,

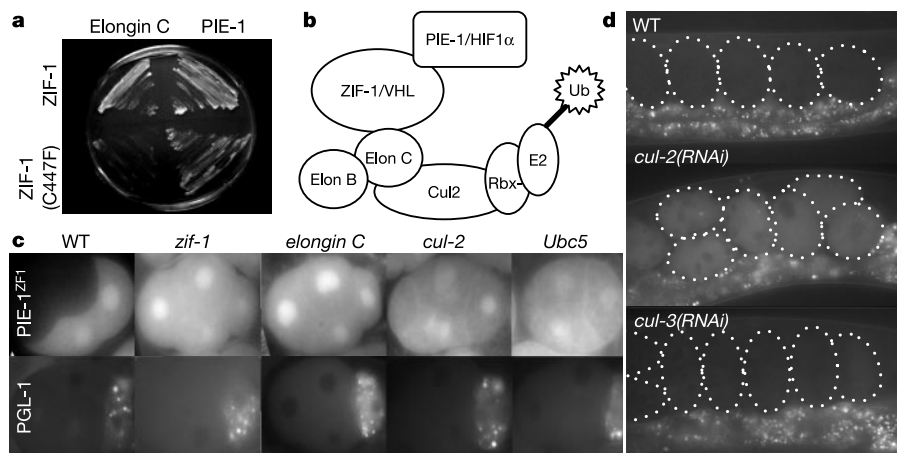


Figure 4 ZIF-1 may function as a substrate-recruitment factor for an elongin C/CUL-2 E3 ubiquitin ligase. **a**, Yeasts expressing ZIF-1, or ZIF-1 with a SOCS mutation, and elongin C or PIE-1 as indicated. Growth indicates interaction. **b**, Diagram after ref. 28 depicting E3 ligase with its associated E2, substrate-recruitment subunits (ZIF-1/VHL) and substrates (PIE-1/HIF1 α). **c**, Four-cell embryos expressing GFP::PIE-1^{ZF1} or GFP::PGL-1 and partially depleted by RNAi for the indicated genes. Visualization of PGL-1

and GFP::PIE-1^{ZF1} in the same *cul-2(RNAi)* embryos confirmed that partial *cul-2* depletion blocks degradation without affecting overall polarity (data not shown). **d**, Live embryos (outlined) expressing GFP::ZIF-1, driven by the *pie-1* promoter (Methods), in wild-type mothers or mothers exposed to *cul-2* or *cul-3* double-stranded RNA. Punctate fluorescence below the embryos is gut autofluorescence from the mother.

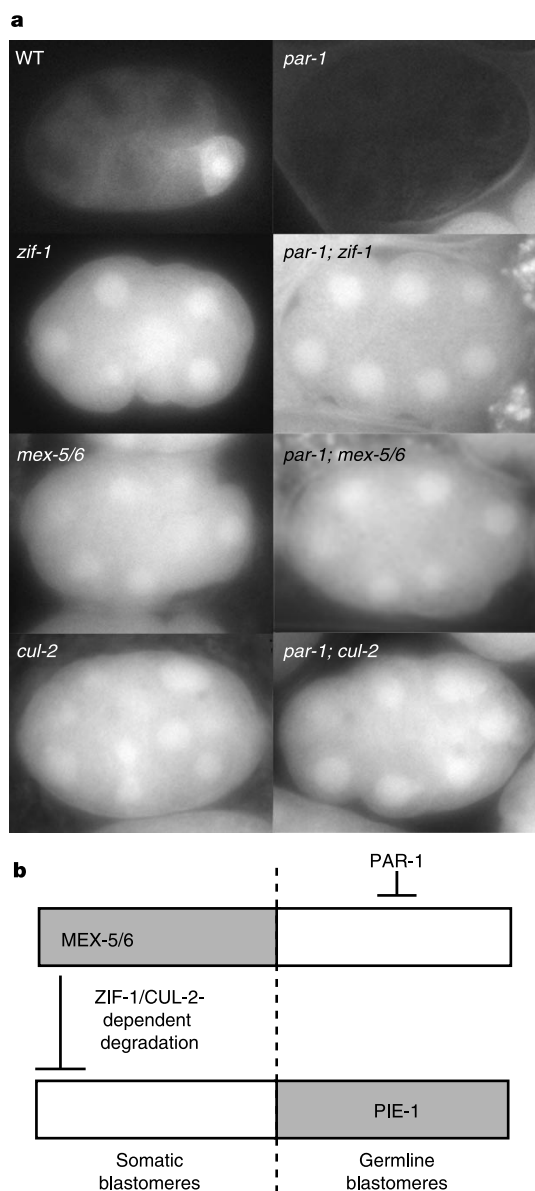


Figure 5 PAR-1 inhibits, and MEX-5 and MEX-6 activate, ZIF-1-dependent degradation. **a**, 12/15-cell embryos expressing GFP::PIE-1^{ZF1} with genes inactivated by RNAi. GFP::PIE-1^{ZF1} is degraded in somatic blastomeres in wild type, and in all cells in *par-1(RNAi)* embryos. Ubiquitous degradation in *par-1(RNAi)* requires ZIF-1, MEX-5, MEX-6 and CUL-2. Similar results were obtained with PIE-1::GFP (Supplementary Information). **b**, Schematic summarizing the epistatic relationships deduced from **a** and ref. 6.

but the mechanism involved in that process is not known, and is distinct from the one described here as it does not require ZIF-1 (Fig. 2g) or PIE-1^{ZF1} (Fig. 2d; see also ref. 10).

Our findings demonstrate that restriction of CCH finger proteins to the germ line requires degradation in somatic lineages. Instability outside of the germ plasm has been observed for other germ plasm components, including P granules in *C. elegans*²⁰, Oskar protein in *Drosophila*²¹ and Vasa protein in zebrafish²². Consistent with our findings, stabilization of Oskar requires PAR-1 (ref. 21). We propose that degradation in the soma coupled with PAR-1-dependent protection in the germ line is a commonly used mechanism to restrict germ plasm components to the germ lineage. Our studies implicate the ECS family of E3 ubiquitin ligases¹³ in the degradation of CCH finger proteins. The recent identification of

Gustavus, a SOCS-box protein required for Vasa localization in *Drosophila*²³, raises the possibility that ECS E3 ligases also regulate other germ plasm components. □

Methods

GFP fusions

GFP fusions were driven by *pie-1* promoter (and 3' UTR)²⁴, which is active during oogenesis (maternal expression). To determine which MEX-5 finger is sufficient for degradation, MEX-5^{ZF1} (QPPNYKTRLCMMHAGIKPCDMGARCKFAHGLKELRAT) and MEX-5^{ZF2} (PNNKYKTKLCKNFARGGTGFCPYGLRCEVHPTDKEFQN) were fused singly to GFP. MEX-5^{ZF2} (ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003), but not MEX-5^{ZF1} (data not shown), was sufficient to target GFP for degradation. Localization dynamics were analysed by time-lapse microscopy as in ref. 9. Movies of protein localization in embryos can be viewed at ftp://ftp.wormbase.org/pub/wormbase/datasets/derenzo_2003.

Yeast two hybrid

PIE-1^{ZF1}–^{ZF2} (amino acids 92–217) or ZIF-1 (F59B2.6) was fused to GAL-4 DNA-binding domain (pAS1-CYH1) and transformed into yeast PJ69-4A. *Caenorhabditis elegans* Gal4 activation domain library (a gift from Z. Zhao and B. Horvitz) was used as prey using standard procedures²⁵. Approximately 600,000 transformants were screened in each screen.

RNAi

RNA interference was performed using the feeding method²⁶. For *zif-1(RNAi)*, *Escherichia coli* expressing *zif-1* double-stranded RNA was plated on IPTG-containing plates for 4 h, before adding L4 hermaphrodites. Embryos were examined 27 h later; 95–100% of embryos had mislocalized PIE-1::GFP and 80–100% did not hatch (embryonic lethality).

Received 30 April; accepted 10 July 2003; doi:10.1038/nature01887.

Published online 23 July 2003.

- Houston, D. W. & King, M. L. Germ plasm and molecular determinants of germ cell fate. *Curr. Top. Dev. Biol.* **50**, 155–181 (2000).
- Kloc, M. *et al.* RNA localization and germ cell determination in *Xenopus*. *Int. Rev. Cytol.* **203**, 63–91 (2001).
- Guedes, S. & Priess, J. R. The *C. elegans* MEX-1 protein is present in germline blastomeres and is a P granule component. *Development* **124**, 731–739 (1997).
- Mello, C. C. *et al.* The PIE-1 protein and germline specification in *C. elegans* embryos. *Nature* **382**, 710–712 (1996).
- Tabara, H., Hill, R. J., Mello, C. C., Priess, J. R. & Kohara, Y. pos-1 encodes a cytoplasmic zinc-finger protein essential for germline specification in *C. elegans*. *Development* **126**, 1–11 (1999).
- Schubert, C. M., Lin, R., de Vries, C. J., Plasterk, R. H. & Priess, J. R. MEX-5 and MEX-6 function to establish soma/germline asymmetry in early *C. elegans* embryos. *Mol. Cell* **5**, 671–682 (2000).
- Blackshear, P. J. Tristetraprolin and other CCCH tandem zinc-finger proteins in the regulation of mRNA turnover. *Biochem. Soc. Trans.* **30**, 945–952 (2002).
- Tenenhaus, C., Schubert, C. & Seydoux, G. Genetic requirements for PIE-1 localization and inhibition of gene expression in the embryonic germ lineage of *Caenorhabditis elegans*. *Dev. Biol.* **200**, 212–224 (1998).
- Cuenca, A. A., Schetter, A., Aceto, D., Kempthues, K. & Seydoux, G. Polarization of the *C. elegans* zygote proceeds via distinct establishment and maintenance phases. *Development* **130**, 1255–1265 (2003).
- Reese, K. J., Dunn, M. A., Waddle, J. A. & Seydoux, G. Asymmetric segregation of PIE-1 in *C. elegans* is mediated by two complementary mechanisms that act through separate PIE-1 protein domains. *Mol. Cell* **6**, 445–455 (2000).
- Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
- Detwiler, M. R., Reuben, M., Li, X., Rogers, E. & Lin, R. Two zinc finger proteins, OMA-1 and OMA-2, are redundantly required for oocyte maturation in *C. elegans*. *Dev. Cell* **1**, 187–199 (2001).
- Kile, B. T. *et al.* The SOCS box: a tale of destruction and degradation. *Trends Biochem. Sci.* **27**, 235–241 (2002).
- Kamura, T. *et al.* Activation of HIF1alpha ubiquitination by a reconstituted von Hippel-Lindau (VHL) tumor suppressor complex. *Proc. Natl Acad. Sci. USA* **97**, 10430–10435 (2000).
- Feng, H. *et al.* CUL-2 is required for the G1-to-S-phase transition and mitotic chromosome condensation in *Caenorhabditis elegans*. *Nature Cell Biol.* **1**, 486–492 (1999).
- Zhou, P. & Howley, P. M. Ubiquitination and degradation of the substrate recognition subunits of SCF ubiquitin-protein ligases. *Mol. Cell* **2**, 571–580 (1998).
- Galan, J. M. & Peter, M. Ubiquitin-dependent degradation of multiple F-box proteins by an autocatalytic mechanism. *Proc. Natl Acad. Sci. USA* **96**, 9124–9129 (1999).
- Wierbelauer, C. *et al.* The F-box protein Skp2 is a ubiquitylation target of a Cul1-based core ubiquitin ligase complex: evidence for a role of Cul1 in the suppression of Skp2 expression in quiescent fibroblasts. *EMBO J.* **19**, 5362–5375 (2000).
- Guo, S. & Kempthues, K. J. par-1, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* **81**, 611–620 (1995).
- Hird, S. N., Paulsen, J. E. & Strome, S. Segregation of germ granules in living *Caenorhabditis elegans* embryos: cell-type-specific mechanisms for cytoplasmic localisation. *Development* **122**, 1303–1312 (1996).
- Riechmann, V., Gutierrez, G. J., Filardo, P., Nebreda, A. R. & Ephrussi, A. Par-1 regulates stability of the posterior determinant Oskar by phosphorylation. *Nature Cell Biol.* **4**, 337–342 (2002).
- Wolke, U., Weidinger, G., Kopranner, M. & Raz, E. Multiple levels of posttranscriptional control lead to germ line-specific gene expression in the zebrafish. *Curr. Biol.* **12**, 289–294 (2002).
- Styhler, S., Nakamura, A. & Lasko, P. VASA localization requires the SPRY-domain and SOCS-box containing protein, GUSTAVUS. *Dev. Cell* **3**, 865–876 (2002).

24. Strome, S. *et al.* Spindle dynamics and the role of gamma-tubulin in early *C. elegans* embryos. *Mol. Biol. Cell* **12**, 1751–1764 (2001).
25. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
26. Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854 (1998).
27. Seydoux, G. & Fire, A. Whole-mount in situ hybridization for the detection of RNA in *Caenorhabditis elegans* embryos. *Methods Cell Biol.* **48**, 323–337 (1995).
28. Ivan, M. *et al.* HIF- α targeted for VHL-mediated destruction by proline hydroxylation: implications for O₂ sensing. *Science* **292**, 464–468 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We dedicate this study to the memory of D. Nathans. We thank E. Kipreos for critical reading of the manuscript; Z. Zhao and B. Horvitz for the yeast two-hybrid library; Y. Kohara for the POS-1 antibody; B. Aman and J. Berg for many discussions about CCH fingers; and A. Cuenca for help with the movies. This work was supported by a NIH grant and by the Steve and Michelle Kirsch Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.S. (gseydoux@jhmi.edu). The sequences of ZIF-1 and elongin C can be retrieved from GenBank with accession numbers CAA77583 and NP_497405, respectively.

RanGTP mediates nuclear pore complex assembly

Tobias C. Walther, Peter Askjaer, Marc Gentzel, Anja Habermann, Gareth Griffiths, Matthias Wilm, Iain W. Mattaj & Martin Hetzer

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

In metazoa, the nuclear envelope breaks down and reforms during each cell cycle. Nuclear pore complexes (NPCs), which serve as channels for transport between the nucleus and cytoplasm¹, assemble into the reforming nuclear envelope in a sequential process involving association of a subset of NPC proteins, nucleoporins, with chromatin followed by the formation of a closed nuclear envelope fenestrated by NPCs^{2–7}. How chromatin recruitment of nucleoporins and NPC assembly are regulated is unknown. Here we demonstrate that RanGTP production is required to dissociate nucleoporins Nup107, Nup153 and Nup358 from Importin β , to target them to chromatin and to induce association between separate NPC subcomplexes. Additionally, either an excess of RanGTP or removal of Importin β induces formation of NPC-containing membrane structures—annulate lamellae—both *in vitro* in the absence of chromatin and *in vivo*. Annulate lamellae formation is strongly and specifically inhibited by an excess of Importin β . The data demonstrate that RanGTP triggers distinct steps of NPC assembly, and suggest a mechanism for the spatial restriction of NPC assembly to the surface of chromatin.

The GTPase Ran drives nucleocytoplasmic transport during interphase and both spindle formation and nuclear envelope (NE) reformation during mitosis^{8–11}. It has also been suggested^{12,13} that Ran functions in NPC assembly, but neither its targets nor its mechanism in this process have been defined. To assess whether Ran might function in nucleoporin recruitment early during NE formation, we incubated sperm chromatin with membrane-free dilute cytosol from *Xenopus* egg extracts and analysed the chromatin using monoclonal Ab414, which recognizes four different nucleoporins¹⁴. Weak, punctate Ab414 staining was detectable on the chromatin surface (Fig. 1A, top panels). Addition of membranes did not increase the Ab414 signal (data not shown), indicating that later steps of NE formation were inefficient in the diluted cytosol.

Addition of 0.1% Triton X-100 to complete extracts abolished membrane association with chromatin but did not prevent accumulation of Ab414 signal on chromatin (data not shown).

Next we added two mutants of Ran, T24N and Q69L, to the reaction. RanT24N inhibits RanGTP production by blocking RCC1, Ran's exchange factor, whereas RanQ69L is incapable of GTP hydrolysis^{15,16}. RanT24N inhibited association of Ab414 nucleoporins with chromatin (Fig. 1a), as did depletion of RCC1 or Ran from the cytosol⁹ (data not shown). RanQ69L, in contrast, caused an increase in Ab414 signal (Fig. 1A). Excess wild-type RanGDP (5 μ M) did not affect nucleoporin recruitment while excess wild-type RanGTP behaved similarly to RanQ69L (Fig. 1A and data not shown).

The chromatin association of nucleoporins Nup107, Nup153 and Nup358, which are recruited to the reforming nucleus early after mitosis, increased in the presence of RanQ69L (Fig. 1B), whereas that of p62, Nup98, Nup214 and Tpr did not change (Fig. 1B and data not shown). In summary, generation of RanGTP by RCC1 appears to be required for association of certain nucleoporins with chromatin during NPC assembly.

NPC assembly during NE formation is restricted to the chromatin surface. If RanGTP is involved in this local restriction, addition of RanGTP throughout the cytoplasm should perturb it. Consistent with previous reports^{8,9}, excess wild-type Ran (green staining, Fig. 1C, upper panels) promoted nuclear assembly and accumulated at the NE while RanQ69L inhibited NE formation (Fig. 1C). Strikingly, RanQ69L concentrated prominently on membrane structures (Fig. 1C, lower panels, green) that contained nucleoporins (see below).

These structures were analysed by transmission electron microscopy (TEM). With wild-type Ran, cisternae and tubules of rough and smooth endoplasmic reticulum (ER) were visible. They occasionally contained NPCs (Fig. 1Da, Dd). Adding 5 μ M RanQ69L resulted in the formation of large membrane stacks several micrometres in length which were fenestrated by pore complexes (Fig. 1Db, De, Df). These structures resembled annulate lamellae (AL)¹⁷, as further verified by immunogold labelling using Ab414 antibody (Fig. 1Dc). Using high-speed extracts no AL were detected in the absence of RanQ69L. AL stacks do form in low-speed extracts¹⁸, but RanQ69L also greatly stimulated this assembly (data not shown).

To quantify RanGTP-induced AL assembly, the presence of nucleoporins in the AL-containing membrane fraction was determined using Ab414. RanQ69L led to a dramatic increase of AL assembly compared to wild-type Ran (Fig. 1E, F), and AL formation was strongly inhibited by RanT24N (Fig. 1F). Subunits of all major nucleoporin sub-complexes tested were present in the AL fraction, while Tpr was not detectable and gp210 levels varied (Supplementary Fig. 1). Quantitative western blotting showed that more than 80% of soluble nucleoporins assembled into AL in the presence of 5 μ M RanQ69L (data not shown).

In *Xenopus* egg extracts individual NPCs assemble into membranes at a low rate, even without chromatin or RanQ69L (Fig. 1Dd), possibly because soluble RCC1 generates a low level of RanGTP which induces NPC insertion. We therefore affinity-depleted RCC1 from cytosol (Fig. 2a, upper panel, some Nup358 was co-depleted, data not shown). In RCC1-depleted extracts the assembly of nucleoporins into membranes was abolished. This effect was specific, because the addition of recombinant RCC1 restored NPC assembly (Fig. 2a, lower panel). In combination with the effect of RanT24N (Fig. 1F), this shows that RanGTP generation by soluble RCC1 is required for NPC assembly into membranes.

Nucleoporins are hyper-phosphorylated during mitosis¹⁹ and their dephosphorylation is thought to be essential for NPC assembly. When mitotic extracts were released into interphase by the addition of Ca²⁺ (ref. 20), nucleoporins were efficiently dephos-

phorylated (Fig. 2b, compare lanes 1, 2 with lanes 3, 4) and assembled into AL (Fig. 2b, lane 7). The addition of RanT24N together with Ca^{2+} blocked AL assembly, demonstrating that nucleoporin dephosphorylation could not induce assembly of NPCs without RanGTP (lanes 4, 8). The effect of RanT24N on pore complex assembly into membranes was reversed by RCC1 (lanes 8, 9). RanQ69L did not induce NPC or AL assembly in mitotic extracts (Fig. 2b, lane 6 and data not shown). Together, these data suggest that both dephosphorylation of nucleoporins and RanGTP are required for post-mitotic NPC assembly.

The NE and AL are specialized regions of the ER. Ran's effect

might therefore be on NPCs directly or on ER membrane assembly. To determine whether Ran is required for ER formation we quantified ER assembly²¹. ER formation was inhibited by cooling on ice but was unaffected by depletion of Ran⁹ (Fig. 3a, upper panels), suggesting that NPC components are the more likely Ran targets.

Mitotic processes are known to be regulated by Ran via transport receptors of the Importin β family and many transport receptor–nucleoporin interactions have been described^{1,22}. Importin β co-immunoprecipitated with nucleoporins Nup107, Nup153 and Nup358 whereas Imp5 and Imp7 did not (Supplementary Fig. 2a).

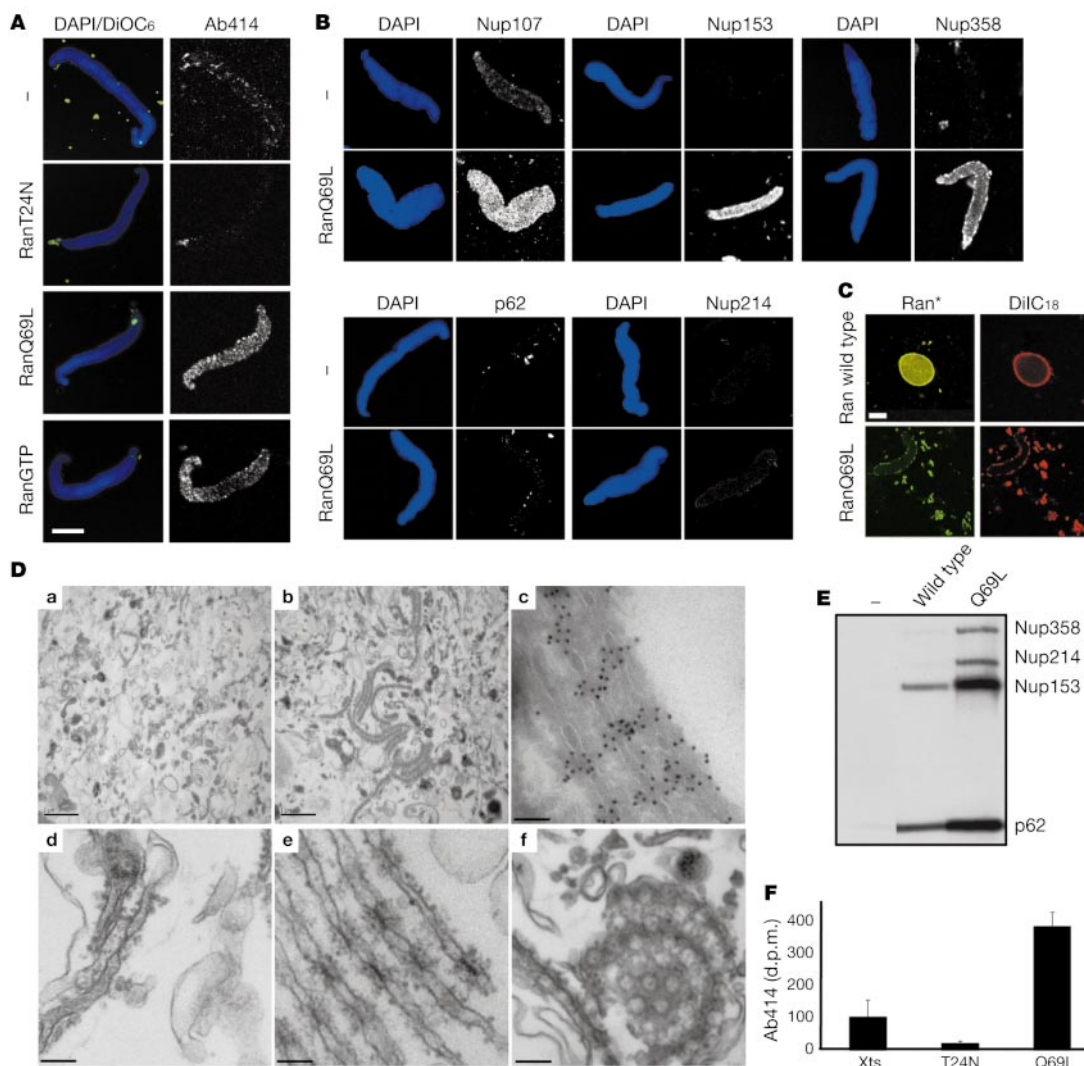


Figure 1 RanGTP stimulates nucleoporin recruitment to sperm chromatin. **A**, Egg extract was diluted 1:1 with buffer S250⁹ and spun at 200,000 g for 60 min to remove membranes. Sperm chromatin was incubated for 60 min with membrane-cleared cytosol with either 5 μM RanT24N, RanQ69L or wild-type RanGTP, re-purified through a 30% sucrose cushion and analysed by immunofluorescence with Ab414 and with DNA stain (DAPI, blue) and membrane staining (DiOC₆, green). Scale bar, 10 μm . **B**, The recruitment of nucleoporins known to associate with chromatin early during nuclear assembly (Nup107, Nup153, Nup358/RanBP2) was stimulated by RanQ69L. Association of the nucleoporins p62 and Nup214/CAN was not affected. **C**, Nuclear assembly with Alexa488-labelled wild-type Ran or RanQ69L (green). Sperm chromatin was incubated with cytosol and DiIC₁₈-labelled membranes (red) together with labelled wild-type Ran (upper panels) or RanQ69L, both indicated as Ran*. The samples were fixed, recovered through a sucrose cushion and analysed by confocal microscopy. Scale bar, 10 μm .

D, Ultrastructure of RanQ69L-induced membrane structures. Reactions were performed as described in **C**, sectioned and analysed by TEM. In the control reactions membrane tubules of ER and vesicles are visible (**Da**). In the presence of 5 μM RanQ69L, stacked membrane sheets are formed (**Db**). The membrane sheets were perforated by NPCs as shown with Ab414 immunogold stain (**Dc**). At higher magnification, membrane fenestration and appearance typical of AL is visible (**De**, **Df**). In the control extract a few individual NPCs can be detected (**Dd**). Scale bars, 2 μm (**Da**, **Db**) and 0.2 μm (**Dc**–**Df**). **E**, Cytosol was incubated for 90 min in the absence (–) or presence of membranes together with wild-type Ran or RanQ69L. Membranes were purified through a sucrose cushion and analysed by western blot with Ab414. **F**, Quantitation of the Ab414 signal from three experiments like that in **E** in which AL was purified from control extracts (Xts) or extracts to which 5 μM RanT24N or RanQ69L were added. d.p.m.: disintegrations per min.

Addition of RanQ69L reduced this co-immunoprecipitation (Supplementary Fig. 2a). Importin β also co-precipitated with p62, Nup153, Nup214 and Nup358, but only in the absence of RanQ69L (Fig. 3b) while the export receptor Crm1 and Nup107 did so only in its presence (Fig. 3b). Importins 5 and 7 did not significantly associate with these nucleoporins.

The effect of RanGTP on cytoplasmic nucleoporin complexes was also studied by gel filtration. Ab414 nucleoporins were present in large complexes (Fig. 3c). Importin β was detectable in complexes across the gradient and RanQ69L efficiently dissociated these complexes, inducing Importin β to move to light fractions. Despite their loss of association with Importin β , complexes containing p62, Nup153 and Nup107—which acts as a marker for an NPC-sub-complex required for NPC assembly^{6,7}—were of higher density in the presence of RanQ69L, while nucleoporins 214 and 358 were little affected (Fig. 3c). This data suggested that RanGTP, by removing Importin β , promotes association between nucleoporins

even in the absence of membranes. Consistent with this possibility, Nup107 was co-precipitated with Ab414 only in the presence of RanGTP (Fig. 3b). Anti-Nup153 antibodies also specifically co-precipitated Nup107, but not p62, Nup214 or Nup358 in the presence of RanQ69L (Fig. 3d, data not shown). Similarly, anti-Nup107 antibodies co-precipitated Nup153 when RanQ69L was present and this precipitation was reduced by excess Importin β , indicating that removal of Importin β from complexes containing these two nucleoporins induces their association (Fig. 3d, lower panels). Thus, in addition to their effects on chromatin association and NPC membrane assembly, Ran and Importin β have effects on the association of nucleoporins with one another.

If the release of Importin β is required for the higher-order assembly of nucleoporins then excess Importin β should inhibit NPC assembly. Indeed, Importin β strongly inhibited AL assembly, while Importin α , Importin 5, Importin 7, Importin 8, Transportin, CAS or Crm1 did not (Figs 3e, g). AL assembly was efficiently blocked by excess of either Importin β or Importin $\beta(\Delta N44)$, a mutant of Importin β that does not bind RanGTP²³. Excess RanQ69L reversed the inhibitory effect of Importin β , but not that of Importin $\beta(\Delta N44)$ (Fig. 3f). Thus, RanGTP allows assembly of nucleoporins into membranes by releasing the inhibitory effect of Importin β . Importantly, excess Importin β had no effect on general ER assembly (Fig. 3a, bottom panels).

If dissociation of Importin β from nucleoporins regulates NPC formation, then its absence should induce NPC assembly into membranes. When Importin β was depleted from egg cytosol (Supplementary Fig. 2b) discrete membrane structures were detected in the assembling ER that stained with Ab414 (Fig. 4a). No such structures were seen in mock-depleted extracts and complementation of the depleted extract with recombinant Importin β inhibited their formation (Fig. 4a). A strong Ab414 signal was also detected on the surface of sperm chromatin in the absence of Importin β . However, the formation of a closed NE was inhibited (Fig. 4b, see refs 10, 11).

To determine whether Importin β has a similar role in NPC formation *in vivo* we examined *C. elegans* embryonic cells. Expression of two import receptors, Importin β and transportin, and two export receptors, CAS and Crm1, was reduced by RNA interference. RNAi against all four receptors resulted in 100% embryonic lethality during gastrulation, demonstrating its effectiveness (Supplementary Fig. 3). Importin β RNA interference resulted in a dramatic increase in cytoplasmic structures which stained with Ab414 (Fig. 4c), Nup107 (data not shown) and GFP-emerin, a trans-membrane protein located in the NE-ER membrane network, suggesting that they are AL (Fig. 4c). RNAi directed against Transportin, CAS or Crm1 did not cause this effect.

Taken together, our results suggest that RanGTP regulates NPC assembly by (1) recruiting a subset of nucleoporins to chromatin, (2) promoting association between nucleoporins and (3) triggering NPC insertion into membranes. During nuclear assembly these steps are coupled and involve interaction of nucleoporins with both chromatin and membranes. The presence of RCC1 and Ran on chromatin during mitosis is thought to localize NE formation to the chromatin surface^{8,9,12,24} and may similarly localize NPC formation.

Importin β interacts with multiple nucleoporins, several in a RanGTP-dependent way. These sites of interaction have been thought to function in NPC translocation during nucleocytoplasmic transport^{1,22}, not in NPC assembly. However, because none of the RanGTP dependent sites of interaction with Importin β have been proven to act during NPC translocation it is even possible, if unlikely, that all RanGTP-dependent sites function only in the regulation of NPC assembly and not during transport. Interestingly, the Importin β mutant I178D, which has low affinity for FG repeats in nucleoporins²⁵, inhibits NPC formation (data not shown), suggesting that strong Importin β –FG repeat interaction may not be involved in NPC assembly.

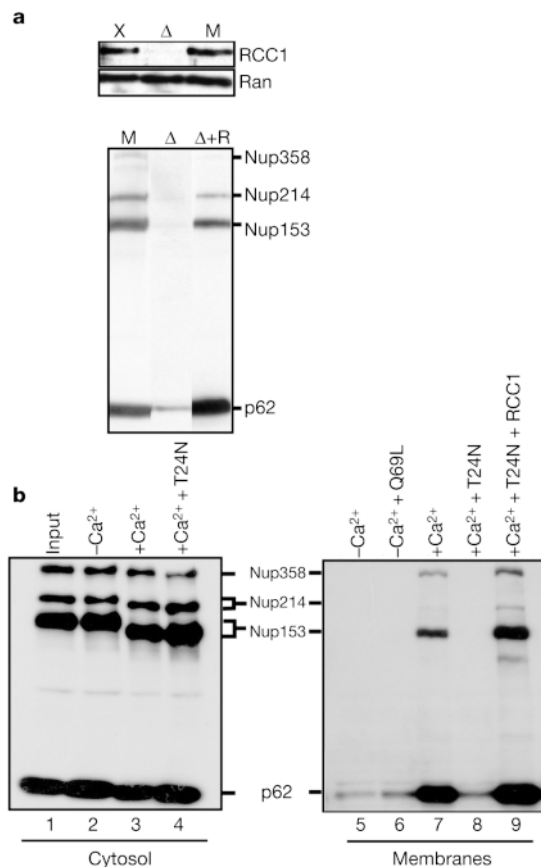


Figure 2 Ran GTP is required for post-mitotic AL assembly. **a**, RCC1 was depleted from cytosol using immobilized RanT24N⁹. The degree of depletion was analysed by western blotting with anti-RCC1 and anti-Ran antibodies: input (X), RCC1-depleted (Δ) and mock-depleted (M). Membranes were incubated with mock or RCC1-depleted cytosol for 90 min in the absence of chromatin, membranes were re-purified through a sucrose cushion and the protein composition analysed by western blotting. Nucleoporin accumulation into membranes was restored by the addition of 0.5 μM RCC1 (Δ + R). **b**, Assembly of NPC-containing membranes in cycling extracts. Mitotic cytosol (lane 1) or cytosol incubated for 90 min without Ca²⁺ (lane 2), with 2 mM Ca²⁺ (lane 3) and with Ca²⁺ together with 5 μM RanT24N (lane 4) was separated by SDS–PAGE and analysed by western blotting with Ab414. Note that dephosphorylation causes a dramatic change in nucleoporin mobility. To analyse AL formation membranes were first incubated with mitotic cytosol in the presence or absence of RanQ69L (lanes 5, 6) or with Ca²⁺-released interphase cytosol (lane 7) treated with RanT24N (lane 8) or RanT24N + RCC1 (lane 9). After repurification, the membrane fractions were separated by SDS–PAGE and analysed with Ab414.

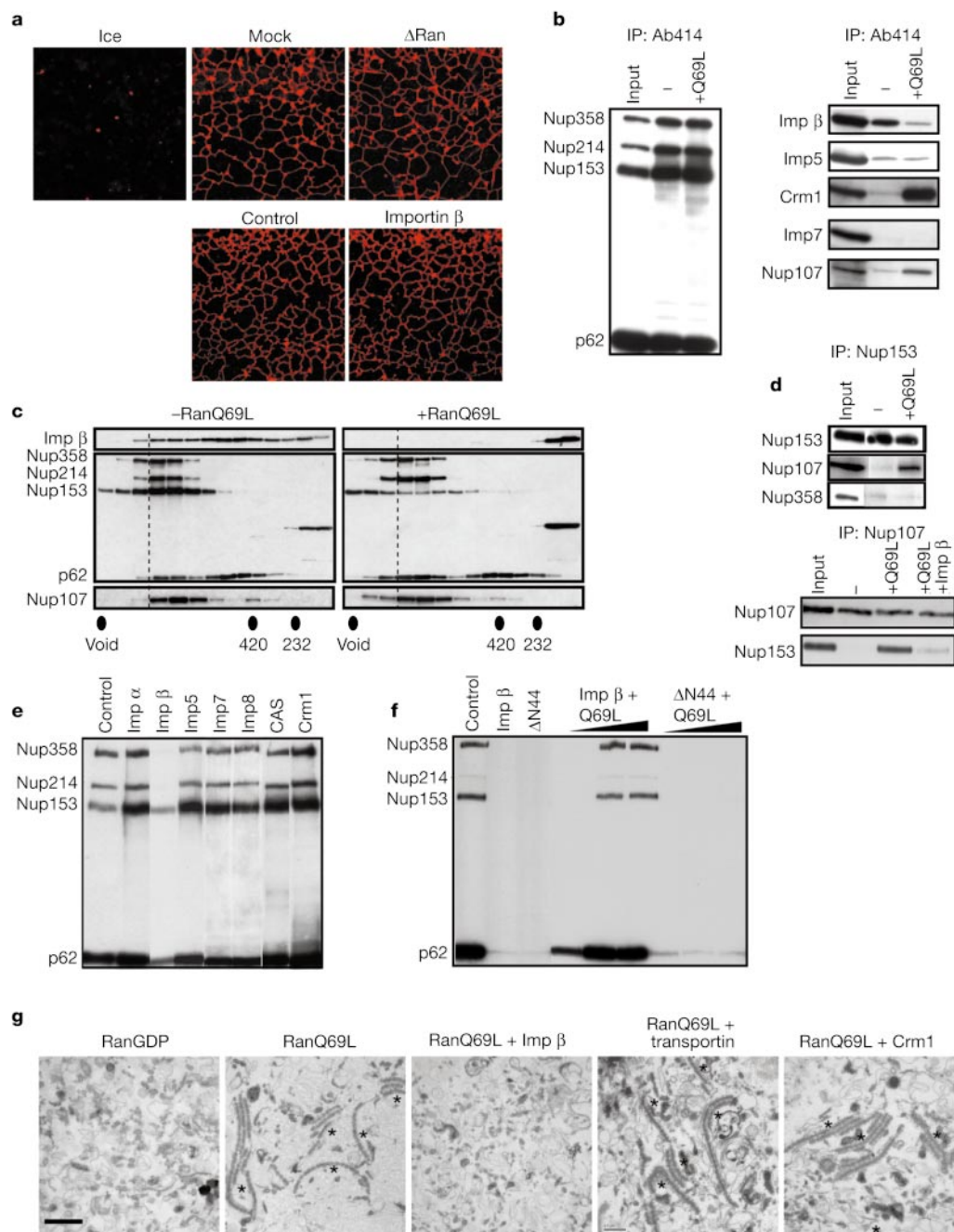


Figure 3 RanGTP-mediated release of Importin β from nucleoporin complexes. **a**, ER assembly was assayed^{5,21} under standard conditions (control), on ice, in mock-depleted or Ran-depleted (Δ Ran) extracts or in the presence of 5 μ M added Importin β . **b**, Nucleoporins were immunoprecipitated from cytosol with Ab414 and their association with transport receptors was analysed by western blotting. Nucleoporins were efficiently precipitated in the presence or absence of RanQ69L (left panel). Importin β associated with Ab414 nucleoporins but Crm1, Imp5, Imp7 and Nup107 did not. The addition of RanQ69L released Importin β from the complexes while Crm1 and Nup107 were recruited (right panel). **c**, Membrane-cleared cytosol was incubated in the absence or presence of RanQ69L and applied to a Superose 6 gel filtration column; fractions were precipitated and analysed by western blotting with Ab414, Nup107 and Importin β antibodies. Mass markers in kDa are indicated. **d**, Extracts were immunoprecipitated with

anti-Nup153 (upper panels) or anti-Nup107 (lower panels) antibodies followed by SDS-PAGE and western blotting with the antibodies indicated on the left of the panels. Reactions were carried out under standard conditions (-) or in the presence of 2 μ M RanQ69L with or without 5 μ M Importin β as indicated. 10% of the input extracts were loaded as a standard. **e**, NPC membrane insertion reactions were performed as described in Fig. 2 but in the presence of 5 μ M Importin α , Importin β , Importin 5, 7, 8, CAS, or Crm1 as indicated. **f**, NPC assembly in the presence of 2 μ M RanQ69L (control) or after adding 5 μ M Importin β or Importin β (Δ N44) in the presence or absence of increasing concentrations of additional RanQ69L (5, 10 and 15 μ M) as indicated. **g**, TEM analysis of AL (asterisks), as in Fig. 1d, in the presence of 5 μ M RanQ69L with or without 10 μ M Importin β , transportin or Crm1 as indicated. Scale bar, 2 μ m.

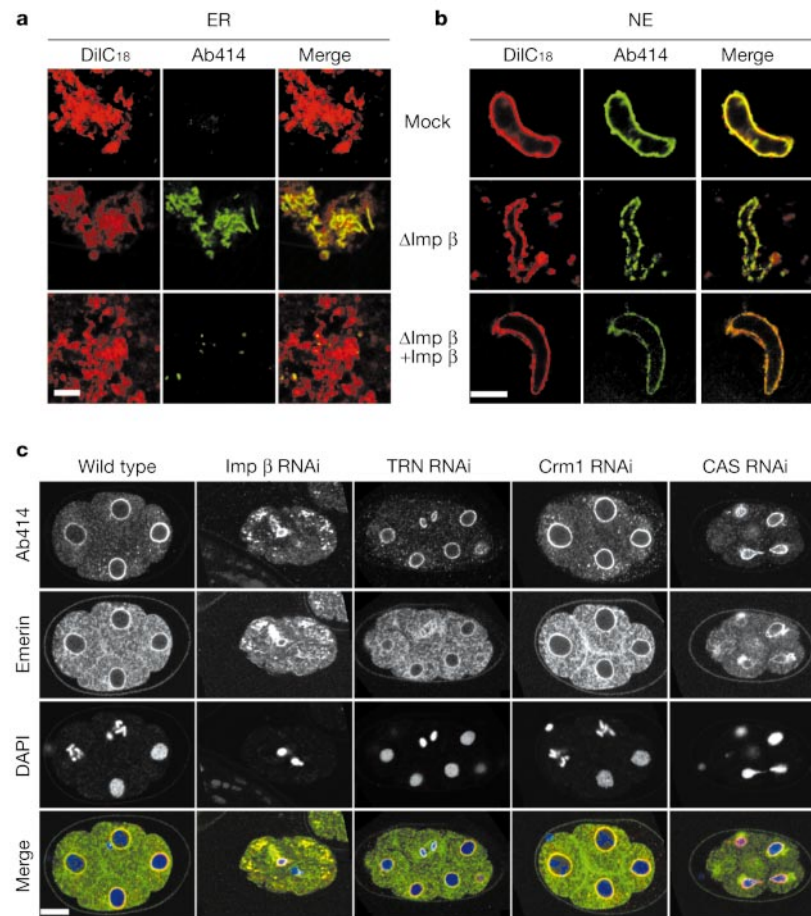


Figure 4 Depletion of Importin β induces the assembly of nucleoporins into membranes. **a, b**, Mock- or Importin β -depleted cytosol was incubated for 30 min with DiIC₁₈-labelled membranes and sperm chromatin in the presence or absence of recombinant Importin β , followed by addition of Ab414/anti-mouse IgG Alexa488 complex and incubation for 10 min. Reactions were fixed with formaldehyde and analysed by confocal microscopy. ER

membranes are shown in **a**, chromatin templates in **b**. Scale bars, 5 μ m, 10 μ m, respectively. **c**, Knock-down of transport receptors using RNAi. Fixed and permeabilized *C. elegans* embryos expressing emerin-GFP were stained for nucleoporins with Ab414 after treatment with dsRNA against the indicated transport receptors. DNA was stained with DAPI dye (blue). Scale bar, 10 μ m.

Previously, NE formation was reported to occur around Sepharose beads coated with Importin β but not with the I178D mutant²⁶ leading to the suggestion that Importin β would have a positive effect in NE assembly, in contrast to the inhibitory role suggested by our data. Importin β tightly associates with several nucleoporins and, since AL form in these extracts in the absence of chromatin, we suggest that Importin β beads may selectively accumulate AL on their surface, which could explain the observations of Zhang *et al.*²⁶. However, since the incorporation of NPCs is not required for closed NE assembly^{6,7}, Importin β -nucleoporin interactions are unlikely to play an essential positive role in NE assembly.

In *S. cerevisiae*, the import receptor Kap121 and RanGTP are required to deliver Nup53p to the NPC²⁷. These data were interpreted in terms of a requirement for nucleocytoplasmic transport in nucleoporin delivery. In combination with our observations, they may, however, suggest that transport receptors function generally as regulators of NPC assembly. RanGTPase is required for correct NE assembly in *Xenopus*^{8,9,12} and in yeasts^{13,28}. It will be interesting to determine whether transport receptors also have a role in NE membrane fusion events.

Finally, our data highlight the double requirement for both RanGTP production and nucleoporin dephosphorylation during NPC assembly. This provides an excellent example of how temporal (phosphorylation state) and spatial (RanGTP production) regulation can be combined to provide order and organization in cells. □

Methods

Recombinant proteins and antibodies

Ran, RanQ69L, RanT24N and the transport receptors Importin β , Importin β (Δ 44), transportin, CAS, Crm1 and Importin 7 and 8 were prepared as previously described^{6,9,23}. Importin 5 was a gift of C. Schatz. Protein-A-tagged IBB and IBBtrunc were expressed as described⁹. Anti-gp62 was a gift from G. Krohne.

Nuclear assembly and related methods

Most methods were previously described^{5,9}. For TEM, samples were embedded in EPON, ultrathin sections were prepared, negatively stained and viewed with a Phillips BioTWIN²⁹. Mitotic extracts were released into interphase as described²⁰.

Biochemical methods

For immunoprecipitation and gel filtration experiments, extract was diluted to 10 mg ml⁻¹ in S250 (ref. 9) supplemented with 150 mM NaCl and centrifuged for 1 h at 200,000 g. Extracts were incubated for 30 min with an energy regenerating system and RanQ69L as indicated, and analysed by gel filtration on Superose 6 (Pharmacia). Samples of 600 μ l were collected, precipitated with 10% TCA and analysed by western blotting. For immunoprecipitations, the indicated antibodies were incubated with 300 μ l extract for 1 h. Protein A-Dynabeads were added and incubation continued for 1 h. After several washes with S250, proteins were eluted with 2% SDS and analysed by western blotting. Depletion of Importin β was performed with zzIBB immobilized on IgG-Sepharose beads and was carried out in the presence of 1 μ M RanGTP to reduce co-depletion of Importin- β -binding proteins. After depletion, 500 nM RanGAP and RanBP1 were added to remove the RanGTP and extracts were frozen. Control extracts were incubated with zzIBBtrunc. Depletion of RCC1 and of Ran was as described⁹.

RNA interference

RNA interference was performed by feeding worms that express GFP::EMR-1 with bacteria that express double-stranded RNA¹¹. For generation of RNAi constructs ~900 nt

sequences were PCR-amplified by using *C. elegans* genomic DNA as template. The DNA fragments were inserted into pPD129.36 (ref. 30). As a negative control in RNAi experiments the empty pPD129.36 vector was used. Sequences of primers are listed in the legend to Supplementary Fig. 3.

Received 14 May; accepted 10 July 2003; doi:10.1038/nature01898.
Published online 30 July 2003.

- Vasu, S. K. & Forbes, D. J. Nuclear pores and nuclear assembly. *Curr. Opin. Cell Biol.* **13**, 363–375 (2001).
- Belgareh, N. *et al.* An evolutionarily conserved NPC subcomplex, which redistributes in part to kinetochores in mammalian cells. *J. Cell Biol.* **154**, 1147–1160 (2001).
- Bodoor, K. *et al.* Sequential recruitment of NPC proteins to the nuclear periphery at the end of mitosis. *J. Cell Sci.* **112**, 2253–2264 (1999).
- Daigle, N. *et al.* Nuclear pore complexes form immobile networks and have a very low turnover in live mammalian cells. *J. Cell Biol.* **154**, 71–84 (2001).
- Hetzler, M. *et al.* Distinct AAA-ATPase p97 complexes function in discrete steps of nuclear assembly. *Nature Cell Biol.* **3**, 1086–1091 (2001).
- Walther, T. C. *et al.* The conserved Nup107–160 complex is critical for nuclear pore complex assembly. *Cell* **113**, 195–206 (2003).
- Harel, A. *et al.* Removal of a single pore subunit results in vertebrate nuclei devoid of nuclear pores. *Mol. Cell* **11**, 853–864 (2003).
- Zhang, C. & Clarke, P. R. Chromatin-independent nuclear envelope assembly induced by Ran GTPase in *Xenopus* egg extracts. *Science* **288**, 1429–1432 (2000).
- Hetzler, M., Bilbao-Cortes, D., Walther, T. C., Gruss, O. J. & Mattaj, J. W. GTP hydrolysis by Ran is required for nuclear envelope assembly. *Mol. Cell* **5**, 1013–1024 (2000).
- Bamba, C., Bobinnec, Y., Fukuda, M. & Nishida, E. The GTPase Ran regulates chromosome positioning and nuclear envelope assembly in vivo. *Curr. Biol.* **12**, 503–507 (2002).
- Askjaer, P., Galy, V., Hannak, E. & Mattaj, J. W. Ran GTPase cycle and importins alpha and beta are essential for spindle formation and nuclear envelope assembly in living *Caenorhabditis elegans* embryos. *Mol. Biol. Cell* **13**, 4355–4370 (2002).
- Zhang, C., Goldberg, M. W., Moore, W. J., Allen, T. D. & Clarke, P. R. Concentration of Ran on chromatin induces decondensation, nuclear envelope formation and nuclear pore complex assembly. *Eur. J. Cell Biol.* **81**, 623–633 (2002).
- Ryan, K. J., McCaffery, J. M. & Went, S. R. The Ran GTPase cycle is required for yeast nuclear pore complex assembly. *J. Cell Biol.* **160**, 1041–1053 (2003).
- Davis, L. I. & Blobel, G. Identification and characterization of a nuclear pore complex protein. *Cell* **45**, 699–709 (1986).
- Bischoff, F. R., Klebe, C., Kretschmer, J., Wittinghofer, A. & Ponstingl, H. RanGAP1 induces GTPase activity of nuclear Ras-related Ran. *Proc. Natl Acad. Sci. USA* **91**, 2587–2591 (1994).
- Klebe, C., Prinz, H., Wittinghofer, A. & Goody, R. S. The kinetic mechanism of Ran-nucleotide exchange catalyzed by RCC1. *Biochemistry* **34**, 12543–12552 (1995).
- Kessel, R. G. Annulate lamellae: a last frontier in cellular organelles. *Int. Rev. Cytol.* **133**, 43–120 (1992).
- Dabauvalle, M. C., Loos, K., Merkert, H. & Scheer, U. Spontaneous assembly of pore complex-containing membranes ("annulate lamellae") in *Xenopus* egg extract in the absence of chromatin. *J. Cell Biol.* **112**, 1073–1082 (1991).
- Macaulay, C., Meier, E. & Forbes, D. J. Differential mitotic phosphorylation of proteins of the nuclear pore complex. *J. Biol. Chem.* **270**, 254–262 (1995).
- Murray, A. in *Cell Cycle Extracts in Xenopus laevis. Practical Uses in Cell and Molecular Biology* (eds Kay, B. K. & Peng, H. B.) 581–605 (Academic, San Diego, 1991).
- Dreier, L. & Rapoport, T. A. In vitro formation of the endoplasmic reticulum occurs independently of microtubules by a controlled fusion reaction. *J. Cell Biol.* **148**, 883–898 (2000).
- Rout, M. P. & Aitchison, J. D. The nuclear pore complex as a transport machine. *J. Biol. Chem.* **276**, 16593–16596 (2001).
- Kutay, U., Izaurralde, E., Bischoff, F. R., Mattaj, J. W. & Gorlich, D. Dominant-negative mutants of importin-beta block multiple pathways of import and export through the nuclear pore complex. *EMBO J.* **16**, 1153–1163 (1997).
- Hinkle, B. *et al.* Chromosomal association of Ran during meiotic and mitotic divisions. *J. Cell Sci.* **115**, 4685–4693 (2002).
- Bayliss, R., Littlewood, T., Strawn, L. A., Went, S. R. & Stewart, M. GLFG and FxGF nucleoporins bind to overlapping sites on importin-beta. *J. Biol. Chem.* **277**, 50597–50606 (2002).
- Zhang, C., Hutchins, J. R., Muhlhauter, P., Kutay, U. & Clarke, P. R. Role of importin-beta in the control of nuclear envelope assembly by Ran. *Curr. Biol.* **12**, 498–502 (2002).
- Lusk, C. P., Makhnevych, T., Marelli, M., Aitchison, J. D. & Wozniak, R. W. Karyopherins in nuclear pore biogenesis: a role for Kap121 in the assembly of Nup53p into nuclear pore complexes. *J. Cell Biol.* **159**, 267–278 (2002).
- Demeter, J., Morpheus, M. & Sazer, S. A mutation in the RCC1-related protein pim1 results in nuclear envelope fragmentation in fission yeast. *Proc. Natl Acad. Sci. USA* **92**, 1436–1440 (1995).
- Griffiths, G. *Fine Structure Immuno-cytochemistry* 137–191 (Springer, Berlin, 1993).
- Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854–857 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Schatz for recombinant RanBP5, V. Galy for experimental advice, G. Krohne for antibodies, K. Ribbeck for discussions and W. Antonin, J. Ellenberg, V. Galy, O. Gruss and E. Izaurralde for critically reading the manuscript. M.H. was supported by the Austrian Programme for Advanced Research and Technology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.W.M. (mattaj@embl-heidelberg.de) or M.H. (hetzler@embl-heidelberg.de).

Phospholipase C γ activates Ras on the Golgi apparatus by means of RasGRP1

Treva G. Bivona¹, Ignacio Pérez de Castro², Ian M. Ahearn¹, Theresa M. Grana³, Vi K. Chiu¹, Peter J. Lockyer⁴, Peter J. Cullen⁵, Angel Pellicer², Adrienne D. Cox³ & Mark R. Phillips¹

¹Departments of Medicine, Cell Biology, Pharmacology and

²Pathology, New York University School of Medicine, 550 First Avenue, New York, New York 10016, USA

³Departments of Radiation Oncology and Pharmacology, University of North Carolina at Chapel Hill School of Medicine, 101 Manning Drive, Chapel Hill, North Carolina 27599, USA

⁴The Babraham Institute, Cambridge, CB2 4AT, UK

⁵Department of Biochemistry, University of Bristol School of Medical Sciences, Bristol, BS8 1TD, UK

Ras proteins regulate cellular growth and differentiation, and are mutated in 30% of cancers. We have shown recently that Ras is activated on and transmits signals from the Golgi apparatus as well as the plasma membrane^{1,2} but the mechanism of compartmentalized signalling was not determined. Here we show that, in response to Src-dependent activation of phospholipase C γ 1, the Ras guanine nucleotide exchange factor RasGRP1 translocated to the Golgi where it activated Ras. Whereas Ca²⁺ positively regulated Ras on the Golgi apparatus through RasGRP1, the same second messenger negatively regulated Ras on the plasma membrane by means of the Ras GTPase-activating protein CAPRI³. Ras activation after T-cell receptor stimulation in Jurkat cells, rich in RasGRP1, was limited to the Golgi apparatus through the action of CAPRI, demonstrating unambiguously a physiological role for Ras on Golgi. Activation of Ras on Golgi also induced differentiation of PC12 cells, transformed fibroblasts and mediated radioresistance. Thus, activation of Ras on Golgi has important biological consequences and proceeds through a pathway distinct from the one that activates Ras on the plasma membrane.

We have used green fluorescent protein (GFP) fused to the Ras-binding domain of Raf1 (GFP-RBD) as a spatio-temporal probe for active Ras in living cells in order to establish that Ha-Ras on Golgi becomes activated in response to growth factors independently of vesicular transport². We therefore explored the possibility that Ca²⁺, a diffusible second messenger that regulates Ras⁴, transmits the signal between plasma membrane and Golgi. In COS-1 cells, epidermal growth factor (EGF) stimulated a relatively rapid (3–10 min) and transient (reversed by 20–40 min) redistribution of GFP-RBD to the plasma membrane followed by a delayed (20 min) and sustained recruitment to the Golgi apparatus (Fig. 1a). In contrast, cells pretreated with either the Ca²⁺ chelator BAPTA-AM, the calcium channel blocker lanthanum⁵ or the phospholipase C (PLC) inhibitor U73122 showed no activation of Ha-Ras on Golgi despite rapid, robust and sustained activation on plasma membrane (Fig. 1a). Under these conditions the distribution of Ha-Ras was unaltered (Fig. 1a). To confirm dependence on PLC- γ 1 we used PLC- γ 1-deficient fibroblasts⁶. GFP-Ha-Ras was localized to the plasma membrane and Golgi in both PLC- γ 1^{+/+} and PLC- γ 1 reconstituted cells. Whereas reconstituted cells showed a pattern of Ras activation similar to COS-1 cells, recruitment of GFP-RBD in PLC- γ 1^{+/+} cells was restricted to plasma membrane and was sustained (Fig. 1b). A similar pattern was observed in fibroblasts deficient in Src, Yes and Fyn (Fig. 1c). The requirements for both PLC- γ 1 and Src were linked by the observation that Src was required for phosphorylation on tyrosine 783 of PLC- γ , a modification

necessary for its activation (Fig. 1d). Thus both activation of Ha-Ras on Golgi and deactivation on plasma membrane are dependent on the elevation of intracellular Ca^{2+} mediated by the Src-dependent activation of PLC- γ 1.

The requirement for PLC- γ in the activation of Ras on Golgi implicated the involvement of the Ras guanine nucleotide releasing protein (RasGRP) family of guanine nucleotide exchange factors (GEFs)^{4,7}. We tagged each of the four RasGRPs with yellow fluorescent protein (YFP) and observed that only YFP-RasGRP1 and

YFP-RasGRP3 had affinity for endomembrane (not shown). We studied RasGRP1 because, unlike RasGRP3, it is Ras-specific and was dynamically recruited to Golgi in COS-1 cells stimulated with EGF (Fig. 2a; see also Supplementary Fig. 1a). In resting PC12 cells endogenous RasGRP1 was cytosolic but translocated to the Golgi in response to phorbol ester (PMA) or nerve growth factor (NGF) (Fig. 2b). The isolated C1 domain of RasGRP1 targeted GFP to the Golgi (not shown), as has been reported for the C1 domain of protein kinase D⁸. Thus, the microenvironment of the Golgi membrane, known to be rich in diacylglycerol (DAG)⁹, conveys specific affinity for a subset of C1 domains.

Lymphocytes are known to be rich in RasGRP1 (ref. 10). In response to stimulation of T-cell receptors (TCRs), RasGRP1-YFP translocated from the cytosol to the Golgi of Jurkat cells where the exchange factor co-localized with Ha-Ras (Fig. 3a). The kinetics of translocation of RasGRP1-YFP to the Golgi in Jurkat cells (<5 min) were severalfold faster than those observed in fibroblasts. Direct activation of RasGRP1 in Jurkat cells with the DAG analogue

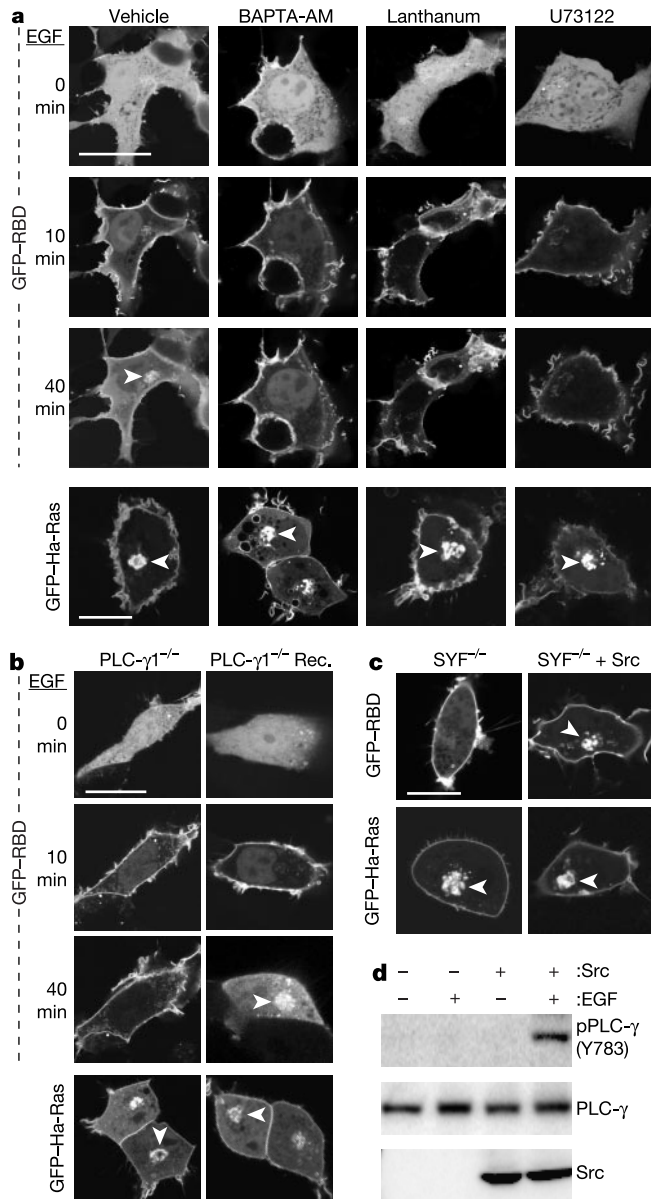


Figure 1 Ha-Ras activation on Golgi and deactivation on plasma membrane requires Src-dependent PLC- γ activity. **a**, COS-1 cells were transfected either with GFP-RBD and Ha-Ras or GFP-Ha-Ras alone (bottom row), serum-starved overnight 24 h after transfection, stimulated with EGF either in the presence of vehicle or drug, as indicated, and imaged alive by LSM. **b**, Fibroblasts either deficient in PLC- γ ^{-/-} (left panels) or reconstituted with (right panels) PLC- γ 1 were transfected, stimulated, and imaged as in **a**. **c**, Fibroblasts deficient in Src, Yes and Fyn (SYF^{-/-}) were co-transfected as in **a** plus either vector (left column) or c-Src (right column), grown in serum overnight, and imaged as in **a**. **d**, SYF^{-/-} cells were transfected with vector or c-Src, serum-starved overnight, then stimulated with vehicle or EGF. Lysates were assayed by immunoblot with anti-phosphorylated PLC- γ (Y783), anti-PLC- γ and anti-Src antiserum. Arrowheads indicate Golgi. Scale bars, 10 μ M.

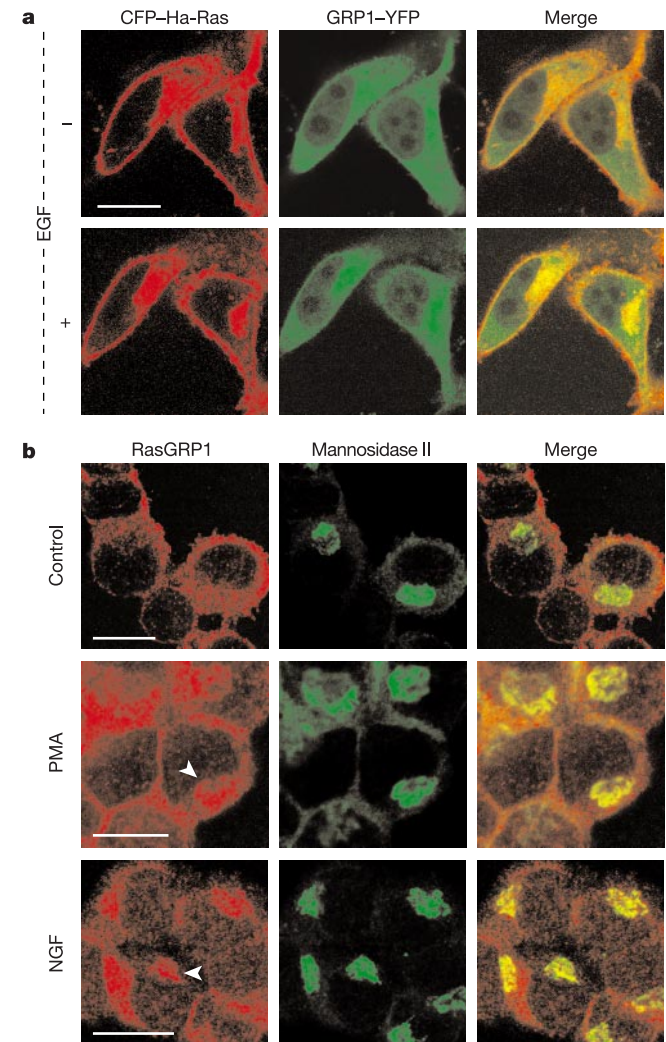


Figure 2 RasGRP1 translocates from cytosol to the Golgi. **a**, COS-1 cells were co-transfected with CFP-Ha-Ras and RasGRP1-YFP, serum-starved overnight 24 h after transfection, and imaged alive by dual-colour LSM before and after stimulation with EGF for 15 min. **b**, PC12 cells treated for 10 min with vehicle (top), PMA (middle) or NGF (bottom) were fixed, permeabilized and stained for endogenous RasGRP1 (red, left panels) and the Golgi marker mannosidase II (green, middle panels). Cells were imaged by LSM using a $\times 100/1.3$ NA objective and a $\times 4$ digital zoom. Arrowheads indicate Golgi; scale bars, 10 μ M.

bryostatin¹¹ resulted in recruitment of GFP-RBD to the Golgi but not the plasma membrane (Fig. 3b). Notably, activation of Jurkat cells by the TCR also induced Ha-Ras activation only on Golgi, a process that was dependent on PLC- γ (Fig. 3c). The paranuclear structure on which cyan fluorescent protein (CFP)-tagged RBD accumulated was identified as Golgi by co-localization with galactosyl transferase-YFP (Fig. 3d). Bystander FRET analysis² confirmed that activation of endogenous Ras in Jurkat cells occurred on the Golgi (Supplementary Fig. 2). Thus, although Ras is localized in T cells on both plasma membrane and Golgi (Fig. 3a), activation occurs only on Golgi. As Ras activation is essential for normal T-cell development and function¹¹, this result establishes unambiguously a physiological role for Ras signalling on Golgi.

To establish that Ras activation on the Golgi apparatus requires RasGRP1 we used dominant-interfering mutants and gene silencing. Overexpression of RasGRP1 comprising a R271E substitution

(Fig. 3e) that abolished GEF catalytic activity¹² or a truncation lacking the GEF domain (not shown) blocked Ras activation in Jurkat cells. In COS-1 cells these mutants blocked Ha-Ras activation at the Golgi but not at the plasma membrane (Supplementary Fig. 1c). Expression from a plasmid¹³ of small interfering (si)RNA targeting RasGRP1 in Jurkat cells, but not control siRNA, silenced RasGRP1 (Supplementary Fig. 3) and blocked TCR-stimulated Ras activation (Fig. 3f). These results are consistent with the observation that Ras activation is abolished in T cells from RasGRP1-deficient mice¹¹.

Overexpression of RasGRP1 in serum-starved COS-1 cells resulted in Ha-Ras activation only on Golgi (Supplementary Fig. 1b). Neither Grb2/SOS (Supplementary Fig. 1b) nor RasGRP2–4 (data not shown) had this effect. To confirm that RasGRP1 activated Ha-Ras *in situ* on Golgi we tested the activity of this GEF towards wild-type Ha-Ras tethered to the Golgi by fusion with KDEL, a transmembrane protein that at steady-state is expressed predominantly on Golgi and not at all on plasma membrane². We observed robust recruitment of GFP-RBD to Golgi (Supplementary Fig. 1b). Thus, whereas knockdown of RasGRP1 inhibits Ha-Ras activation on Golgi, overexpression of this exchange factor has the opposite effect. The ability of overexpressed RasGRP1 to activate Ha-Ras in serum-starved cells is consistent with the ability of RasGRP1 to function as an oncogene^{14,15}, and suggests that RasGRP1 transforms cells by means of signals originating on the Golgi apparatus.

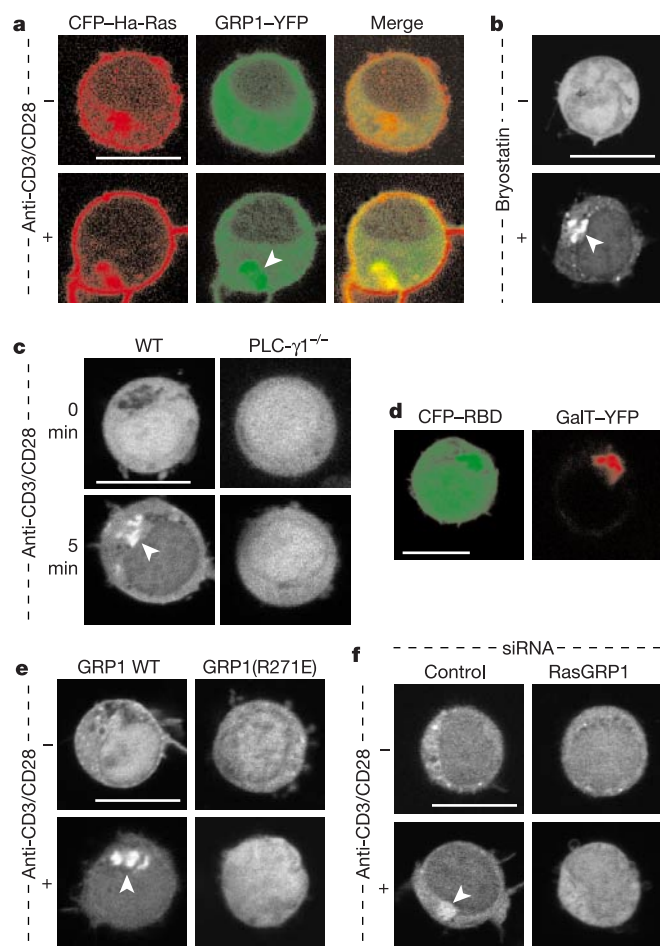


Figure 3 Ras activation in T cells is restricted to Golgi and is RasGRP1-dependent. **a**, Jurkat T cells were co-transfected with CFP-Ha-Ras and RasGRP1-YFP, serum-starved for 2 h beginning 48 h after transfection, and then imaged alive before and after stimulation with anti-CD3 and anti-CD28 antibodies for 5 min. **b**, Jurkat T cells were co-transfected with GFP-RBD and Ha-Ras, and imaged as in **a** with or without bryostatin (5 min). **c**, Jurkat T cells either wild type (WT) or deficient in PLC- γ 1 were co-transfected as in **b** and treated and imaged as in **a**. **d**, Jurkat T cells were co-transfected with Ha-Ras, CFP-RBD and the Golgi marker galactosyltransferase-YFP (GalT-YFP), then treated and imaged as in **a**. **e**, Jurkat T cells were co-transfected with GFP-RBD, Ha-Ras and wild-type RasGRP1 or RasGRP1(R271E), as indicated, and then treated and imaged as in **a**. **f**, Jurkat T cells were co-transfected as in **b** along with either pSuper-scrambled control or pSuper-RasGRP1 siRNA-generating plasmids, and treated and imaged as in **a**. Arrowheads indicate Golgi; scale bars, 10 μ M.

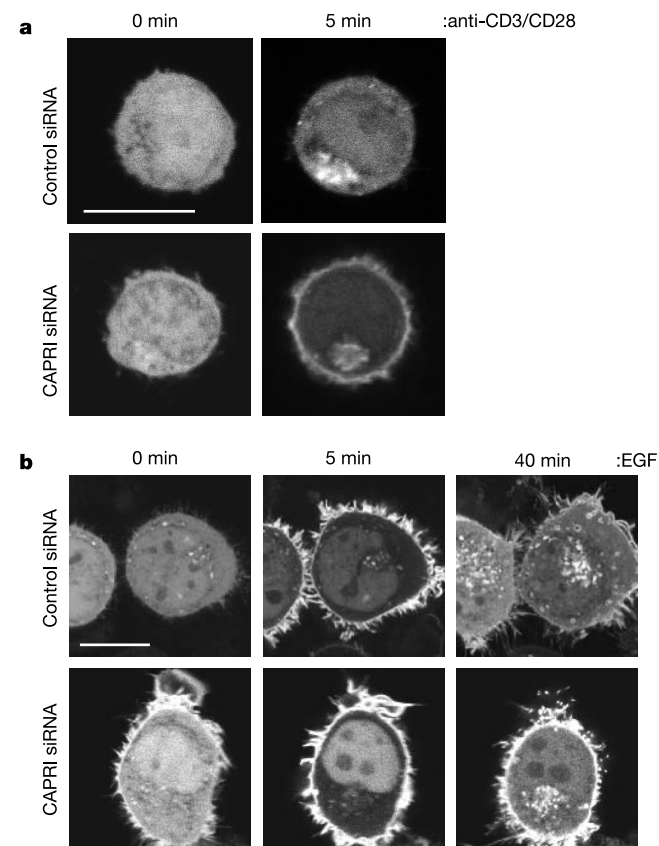


Figure 4 CAPRI deactivates Ras on the plasma membrane. **a**, Jurkat T cells were co-transfected with GFP-RBD and Ha-Ras along with either pSuper-scrambled control or pSuper-CAPRI siRNA-generating plasmids, as indicated, serum-starved for 2 h beginning 48 h after transfection, and imaged alive before and 5 min after stimulation with anti-CD3 and anti-CD28 antibodies. **b**, HeLa cells were co-transfected as in **a**, serum-starved overnight 24 h later, and imaged alive at the indicated times after stimulation with EGF. Scale bars, 10 μ M.

The observation that Ha-Ras was expressed but not activated on the plasma membrane of Jurkat cells (Fig. 3) suggested that Ha-Ras was inhibited on this compartment. Because CAPRI is a Ca^{2+} -sensitive Ras GTPase-activating protein that can translocate to the plasma membrane³, we tested whether this negative regulator of Ras is responsible for blocking Ras activation on the plasma membrane of Jurkat cells. CAPRI siRNA inhibited by >90% the expression of YFP-CAPRI in Jurkat cells (Supplementary Fig. 3c). Notably, whereas Jurkat T cells transfected with control siRNA showed recruitment of GFP-RBD only to Golgi on TCR stimulation, cells transfected with CAPRI siRNA showed robust recruitment to both Golgi and plasma membrane (Fig. 4a). Thus, the suppression of Ras activation on the plasma membrane of Jurkat cells was mediated by CAPRI. We next investigated whether the activity of CAPRI could account for the transient nature of plasma membrane activation in non-lymphoid cells. Whereas HeLa cells transfected with Ha-Ras, GFP-RBD and control siRNA responded in a similar manner to COS-1 cells on stimulation with EGF—that is, early and transient activation of Ha-Ras on plasma membrane followed by sustained activation on Golgi—HeLa cells transfected with CAPRI siRNA showed sustained activation on plasma membrane (Fig. 4b). To determine whether overexpression of CAPRI could specifically suppress plasma membrane Ha-Ras activation in non-lymphoid cells we transfected PC12 cells with GFP-RBD, Ha-Ras and either vector or CAPRI. Whereas Ha-Ras activation on Golgi of CAPRI-overexpressing cells was unaffected, activation on plasma membrane was inhibited (Fig. 5b). Similar results were observed in COS-1 cells (data not shown). Thus, CAPRI acts at the plasma

membrane to reverse Ras activation in non-lymphoid cells and blocks entirely Ras activation on the plasma membrane of T cells. Whereas Ca^{2+} fluxes are relatively short-lived in fibroblasts stimulated with EGF¹⁶, they are sustained in Jurkat T cells stimulated through TCRs¹⁷, suggesting an explanation for the more potent inhibitory effect of CAPRI in Jurkat T cells.

To investigate the physiological role of Ras signalling on the Golgi apparatus we studied PC12 cells. The Ras pathway leads to both proliferation and differentiation of PC12 cells, and it is thought to be the duration of mitogen-activated protein kinase (MAPK) signalling that determines which outcome prevails, with sustained MAPK activation specifying differentiation¹⁸. Thus, sustained activation of Ha-Ras on the Golgi apparatus relative to that on the plasma membrane may have a role in the differentiation pathway. Whereas Ras activation on plasma membrane but not Golgi was abolished in PC12 cells overexpressing CAPRI, neurite outgrowth remained qualitatively (Fig. 5a) and quantitatively (Fig. 5b) intact. Furthermore, endomembrane-restricted, Golgi-enriched KDELR-Ha-Ras61L was more potent in its ability to stimulate neurite outgrowth than endoplasmic-reticulum-enriched, palmitoylation-deficient Ha-Ras61L,181,184S, and was equivalent to natively targeted Ha-Ras61L (Fig. 5c). Thus activation of Ras on Golgi is sufficient to induce PC12 cell differentiation.

To confirm a role for RasGRP1 in PC12 cell differentiation mediated by Golgi-associated Ras, we studied the effects of overexpressing this exchange factor or silencing its gene. Whereas an activated form of RasGRP1 was reported to stimulate neurite outgrowth in PC12 cells¹⁹, overexpressing wild-type RasGRP1 failed

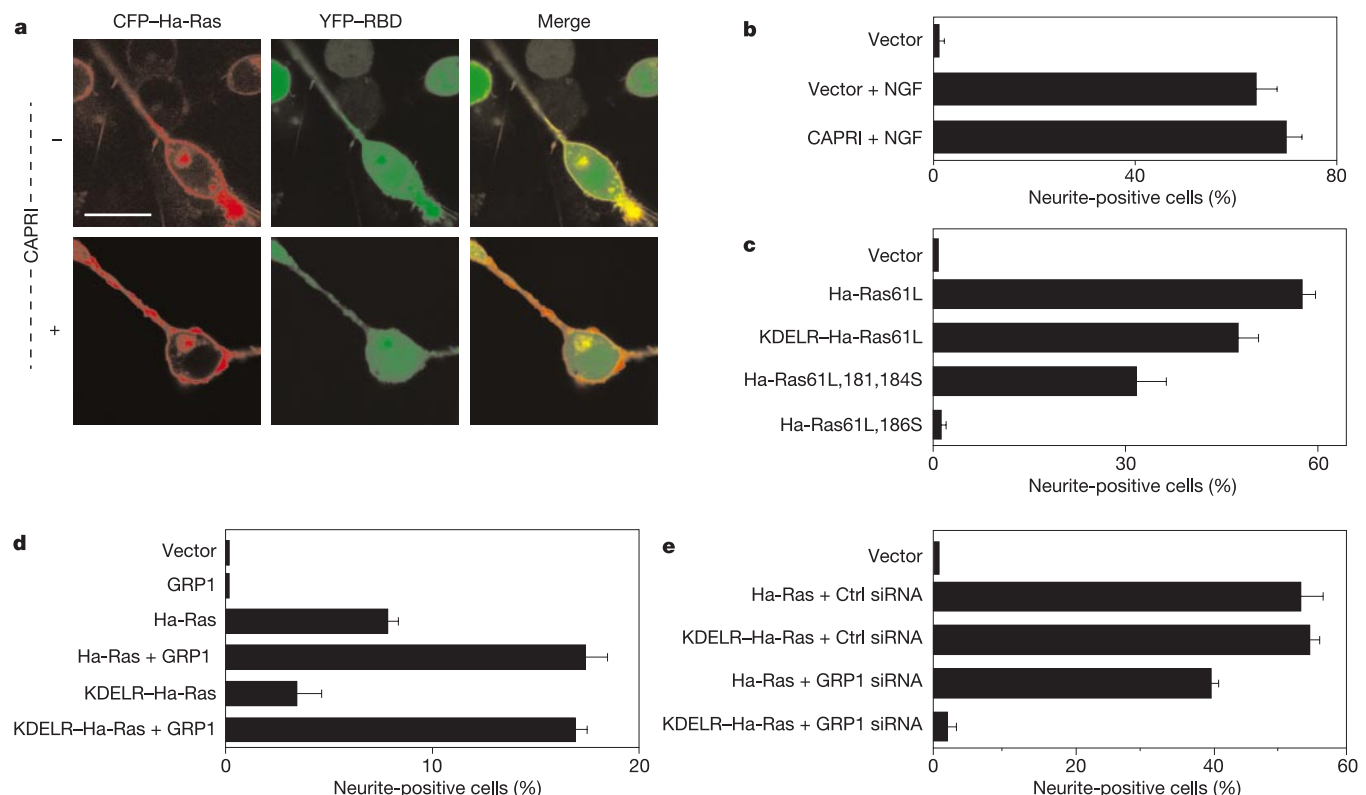


Figure 5 Activation of Ras on Golgi promotes PC12 neuronal differentiation. **a**, PC12 cells were co-transfected with CFP-Ha-Ras, YFP-RBD and either excess vector or CAPRI, and cells exhibiting neurite outgrowth were imaged 48 h later. Scale bars, 10 μm . **b**, PC12 cells were co-transfected with pEGFP and either excess vector or CAPRI, and were stimulated with NGF. Neurite outgrowth was scored at 72 h as described in Methods. **c**, PC12 cells were co-transfected with pEGFP and excess plasmid DNA encoding the Ras

construct indicated, and neurite outgrowth after 48 h was scored as in **b**. **d**, PC12 cells were co-transfected with pEGFP and excess RasGRP1 or vector and excess wild-type Ha-Ras natively targeted or tethered to KDELR. Neurite outgrowth was scored as in **b**. **e**, PC12 cells were co-transfected with pEGFP plus excess wild-type Ha-Ras natively targeted or fused to KDELR and pSuper vectors encoding either control or RasGRP1 siRNA. Neurite outgrowth after 72 h was scored as in **b**.

to induce any neurites (Fig. 5d). However, RasGRP1 synergized with Ha-Ras to induce neurites. Importantly, RasGRP1 synergized equally well with both natively targeted Ha-Ras and Golgi-restricted KDELR–Ha-Ras (Fig. 5d), indicating that the full effect of co-expression of Ha-Ras and the exchange factor could be achieved with Ha-Ras restricted to the endomembrane. Conversely, RasGRP1 siRNA inhibited Ha-Ras-induced neurite outgrowth and abolished KDELR–Ha-Ras-induced differentiation (Fig. 5e). This result indicates that Ha-Ras signalling from the Golgi apparatus in PC12 cells is dependent on endogenous RasGRP1.

To establish further the biological relevance of Ras activation on Golgi we studied two additional readouts of Ras function in response to endomembrane-restricted, activated Ras. Stable expression of KDELR–Ha-Ras61L transformed murine fibroblasts and promoted the acquisition of resistance to ionizing radiation in RIE-1 cells (Supplementary Fig. 4). Thus, three independent measures of Ras function confirmed the biological activity of Golgi-associated Ras.

Our data define a new pathway by which engagement of growth factor receptors or TCRs leads to activation of Ha-Ras on Golgi. This Src/PLC- γ 1/RasGRP1-dependent pathway is distinct from the protein tyrosine kinase receptor/Shc (present or absent)/Grb2/SOS pathway that activates Ras on the plasma membrane. Moreover, by activating both RasGRP1 and the Ras GTPase-activating protein CAPRI, Ca²⁺ can regulate Ras in opposite directions on different subcellular compartments. The binary GTP/GDP switch that constitutes Ras as a signalling element cannot explain the variety of outcomes of Ras activation. By expressing Ras on different subcellular compartments and by using distinct modes of regulation, the cell gains increased capacity to modulate its output. □

Methods

Plasmids

The full or partial coding sequences of the relevant human complementary DNAs were amplified by polymerase chain reaction (PCR) and cloned in-frame (primer sequences and restriction sites available on request) into pEGFP-N1, pEGFP-C3, pECFP-C1, pYFP-N1 (Clontech), pcDNA3.1(+)/Neo (Invitrogen) or pBabe/puro, as indicated. Generation of pEGFP-RBD (GFP-RBD)², pcDNA-KDELR–Ha-Ras³ and PCI-Neo-CAPRI³ was described previously. Plasmids encoding human Grb2 and SOS were provided by E. Skolnik and D. Bar Sagi, respectively. RasGRP1 cDNA was provided by J. Stone. Dominant-interfering RasGRP1(R271E) was generated by site-directed mutagenesis using the Quickchange XL Kit (Stratagene). Dominant-negative RasGRP1(EF-C1) was generated by PCR amplification of codons 472–595. pSuper plasmids encoding templates for siRNAs for RasGRP1, CAPRI or scrambled controls were constructed as described¹³. We verified all plasmid constructs by bi-directional sequencing.

Cell culture and transfection

COS-1 cells, immortalized fibroblasts deficient in Src, Yes and Fyn (SYF^{−/−}) (American Type Culture Collection) or in PLC- γ 1 (PLC- γ 1^{−/−}), or the same cells reconstituted with PLC- γ 1 (gifts of G. Carpenter), were maintained in 5% CO₂ in DMEM medium supplemented with 4 mM L-glutamine, 1.5 g l^{−1} sodium bicarbonate, 4.5 g l^{−1} glucose, and 10% fetal bovine serum (Colorado Serum Co.). Jurkat cells expressing or deficient in PLC- γ 1 were obtained from American Type Culture Collection and were maintained in 5% CO₂ in RPMI medium containing 10% fetal bovine serum. Cells to be examined by fluorescence microscopy were plated at 2 × 10⁵ per plate into 35-mm dishes containing a glass coverslip-covered 15-mm cut out (MatTek) and transfected the next day using SuperFect (Qiagen; COS-1), DMR1 (Invitrogen; Jurkat), or lipofectamine 2000 (Invitrogen; immortalized fibroblasts, PC12) according to the manufacturer's instructions. To ensure co-transfection of untagged cDNAs and fluorescent constructs, a DNA ratio of at least 5:1 (untagged:fluorescent construct) was used. Expression of siRNAs for RasGRP1 or CAPRI was accomplished by co-transfection with tenfold excess of each pSuper construct.

Cell stimulation and imaging

Cells were stimulated, while continuously observing selected cells, by adding 40 ng ml^{−1} EGF, 5 ng ml^{−1} anti-CD3 plus anti-CD28 antibodies, 100 ng ml^{−1} NGF, 100 nM PMA or 100 nM bryostatin to 35-mm plates (MatTek) maintained at 37 °C using a PDMI-2 microincubator (Harvard Apparatus). Treatment of cells with 10 μ M U73122 (Calbiochem), 10 μ M BAPTA-AM (Molecular Probes) or 20 μ M lanthanum was initiated 30 min before stimulations. Living cells were imaged with a Zeiss 510 inverted laser scanning confocal microscope (LSM). A minimum of five 0.45- μ m Z slices were acquired for each cell at each time point, and representative images were chosen to display both plasma membrane and Golgi. Statistical validation that each fluorescence pattern selected

for display is representative is presented in Supplementary Table 1. Tiff images were processed with Adobe Photoshop 7.0.

Indirect immunofluorescence

PC12 cells (10⁵) were plated onto glass coverslips, stimulated with vehicle, PMA or NGF, fixed with 4% paraformaldehyde, and permeabilized and blocked with 0.5% Triton-X100/5% BSA in PBS. The cells were dual-stained with a monoclonal antibody to rat RasGRP1 (m199, Santa Cruz) and a polyclonal antiserum against rat mannosidase II (gift of K. Moreman) (each diluted 1:250), followed by Texas-red-conjugated horse anti-mouse combined with fluorescein isothiocyanate-conjugated goat anti-rabbit antiserum (Jackson ImmunoResearch). Control staining was performed using only the secondary antiserum. Coverslips were mounted with photobleach-retardant medium (DAKO Corp.) and the cells were imaged with a Zeiss 510 inverted LSM.

Immunoblots

SYF^{−/−} or c-Src reconstituted immortalized fibroblasts (10⁶) were plated in 10-cm dishes and 24 h later serum-starved overnight, then stimulated with EGF. Cells were lysed in RIPA buffer containing protease inhibitors, 20 mM NaF and 10 mM NaVO₄. Clarified lysates were analysed by immunoblot with an antibody directed to PLC- γ phosphorylated on tyrosine 783 (Cell Signaling Tech.), PLC- γ (Cell Signaling Tech.) or an antibody against Src (Santa Cruz Biotech.). In all immunoblot analyses, ¹²⁵I-labelled protein A (Amersham) was used as a secondary reagent permitting analysis by phosphorimager.

PC12 cell differentiation

PC12 cells were cultured in DMEM supplemented with 10% fetal bovine serum and 10% horse serum (Colorado Serum Co.), and co-transfected with pEGFP and fivefold excess of the plasmids encoding the indicated untagged proteins using lipofectamine 2000 (Invitrogen). Neurite outgrowth was measured 48 or 72 h later by LSM and quantified as the number of transfected cells (scored by expression of GFP) with extensions longer than the greatest diameter of the cell body. Three fields were examined on duplicate plates.

Received 10 May; accepted 9 June 2003; doi:10.1038/nature01806.

Published online 29 June 2003.

- Choy, E. *et al.* Endomembrane trafficking of ras: the CAAX motif targets proteins to the ER and Golgi. *Cell* **98**, 69–80 (1999).
- Chiu, V. K. *et al.* Ras signalling on the endoplasmic reticulum and the Golgi. *Nature Cell Biol.* **4**, 343–350 (2002).
- Lockyer, P. J., Kupzig, S. & Cullen, P. J. CAPRI regulates Ca(2+) -dependent inactivation of the Ras-MAPK pathway. *Curr. Biol.* **11**, 981–986 (2001).
- Cullen, P. J. & Lockyer, P. J. Integration of calcium and Ras signalling. *Nature Rev. Mol. Cell Biol.* **3**, 339–348 (2002).
- Yue, L., Peng, J. B., Hediger, M. A. & Clapham, D. E. CaT1 manifests the pore properties of the calcium-release-activated calcium channel. *Nature* **410**, 705–709 (2001).
- Ji, Q. S., Ermini, S., Baulida, J., Sun, F. L. & Carpenter, G. Epidermal growth factor signaling and mitogenesis in Plc γ 1 null mouse embryonic fibroblasts. *Mol. Biol. Cell* **9**, 749–757 (1998).
- Ehrhardt, A., Ehrhardt, G. R., Guo, X. & Schrader, J. W. Ras and relatives—job sharing and networking keep an old family together. *Exp. Hematol.* **30**, 1089–1106 (2002).
- Baron, C. L. & Malhotra, V. Role of diacylglycerol in PKD recruitment to the TGN and protein transport to the plasma membrane. *Science* **295**, 325–328 (2002).
- Bankaitis, V. A. Slick recruitment to the Golgi. *Science* **295**, 290–291 (2002).
- Ebinu, J. O. *et al.* RasGRP links T-cell receptor signaling to Ras. *Blood* **95**, 3199–3203 (2000).
- Dower, N. A. *et al.* RasGRP is essential for mouse thymocyte differentiation and TCR signaling. *Nature Immunol.* **1**, 317–321 (2000).
- Park, W., Mosteller, R. D. & Broek, D. Identification of a dominant-negative mutation in the yeast CDC25 guanine nucleotide exchange factor for Ras. *Oncogene* **14**, 831–836 (1997).
- Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
- Ebinu, J. O. *et al.* RasGRP, a Ras guanyl nucleotide-releasing protein with calcium- and diacylglycerol-binding motifs. *Science* **280**, 1082–1086 (1998).
- Tognon, C. E. *et al.* Regulation of RasGRP via a phorbol ester-responsive C1 domain. *Mol. Cell Biol.* **18**, 6995–7008 (1998).
- Peppelenbosch, M. P. *et al.* Rac-dependent and -independent pathways mediate growth factor-induced Ca²⁺ influx. *J. Biol. Chem.* **271**, 7883–7886 (1996).
- Lewis, R. S. Calcium signaling mechanisms in T lymphocytes. *Annu. Rev. Immunol.* **19**, 497–521 (2001).
- Marshall, C. J. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* **80**, 179–185 (1995).
- Yamashita, S. *et al.* CalDAG-GEFIII activation of Ras, R-ras, and Rap1. *J. Biol. Chem.* **275**, 25488–25493 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Stone for anti-RasGRP1 antibodies, G. Carpenter for PLC- γ ^{−/−} cells and E. Skolnik and B. Neel for critically reading the manuscript. This work was supported by grants from the National Institutes of Health (to M.R.P., A.P. and A.D.C.), the New York State Breast Cancer Research Program and the Burroughs Wellcome Fund (to M.R.P.), and the Lister Institute (to P.J.C.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.R.P. (philim01@med.nyu.edu).

Crystal structure of the transfer-RNA domain of transfer-messenger RNA in complex with SmpB

Sascha Gutmann¹, Peter W. Haeberl¹, Laurent Metzinger², Markus Sutter¹, Brice Felden² & Nenad Ban¹

¹Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule Hönggerberg (ETH Zürich), HPK Gebäude, CH-8093 Zürich, Switzerland

²Biochimie Pharmaceutique UPRES JE2311, Université de Rennes I, avenue du Pr. Léon Bernard, CS 34317, F-35043 Rennes Cedex, France

Accurate translation of genetic information into protein sequence depends on complete messenger RNA molecules. Truncated mRNAs cause synthesis of defective proteins, and arrest ribosomes at the end of their incomplete message. In bacteria, a hybrid RNA molecule that combines the functions of both transfer and messenger RNAs (called tmRNA) rescues stalled ribosomes, and targets aberrant, partially synthesized, proteins for proteolytic degradation^{1,2}. Here we report the 3.2-Å-resolution structure of the tRNA-like domain of tmRNA (tmRNA_Δ) in complex with small protein B (SmpB), a protein essential for

biological functions of tmRNA. We find that the flexible RNA molecule adopts an open L-shaped conformation and SmpB binds to its elbow region, stabilizing the single-stranded D-loop in an extended conformation. The most striking feature of the structure of tmRNA_Δ is a 90° rotation of the TΨC-arm around the helical axis. Owing to this unusual conformation, the SmpB–tmRNA_Δ complex positioned into the A-site of the ribosome orients SmpB towards the small ribosomal subunit, and directs tmRNA towards the elongation-factor binding region of the ribosome. On the basis of this structure, we propose a model for the binding of tmRNA on the ribosome.

tmRNA, also referred to as SsrA and 10S RNA, is a highly structured RNA that has properties of both tRNA and mRNA. tmRNA in complex with SmpB³ is a substrate for the same enzymes that process canonical tRNAs. Alanyl-tRNA synthetase (AlaRS) charges tmRNA with alanine, and elongation factor Tu (EF-Tu) delivers alanyl-tmRNA to stalled ribosomes⁴. The incompletely synthesized protein is transferred from the P-site tRNA to the alanyl-tmRNA bound in the A-site of the large ribosomal subunit. The ribosome then—in a process called trans-translation—resumes translation, switching from the aberrant mRNA to the open reading frame encoded by the mRNA region of tmRNA. Ten highly conserved amino acids are added to the incompletely synthesized protein as a degradation tag² before a stop codon at the 3' end of the tmRNA open reading frame terminates translation with the

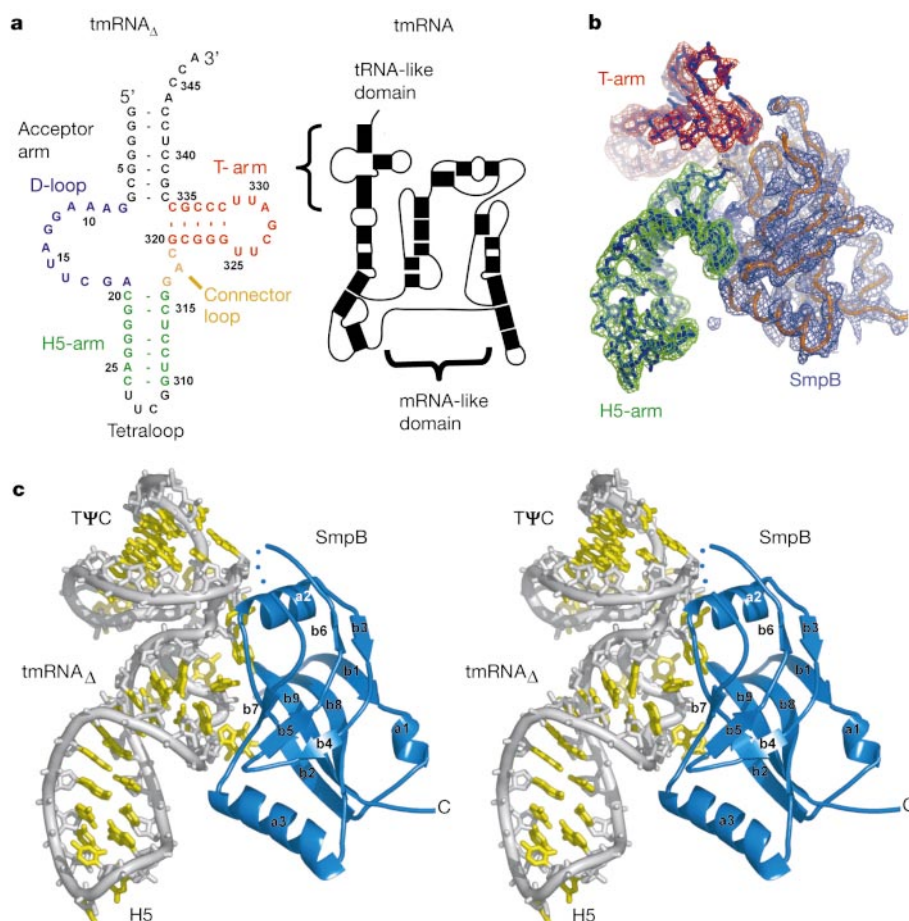


Figure 1 Overview of the SmpB–tmRNA_Δ structure. **a**, Secondary structure diagram of tmRNA and the tmRNA_Δ construct (enlarged) used in this study. The D-arm lacks secondary structure elements. tmRNA_Δ domains are colour-coded: D-loop (blue), H5-arm (green), T-arm (red), connector loop (wheat), acceptor arm and tetraloop (black). **b**, Section of the 2F_o - F_c electron density map contoured at 1σ defining the

SmpB–tmRNA_Δ complex. The map section was colour-coded as in **a** according to the different domains of tmRNA_Δ. Electron density for SmpB is shown in blue and the protein backbone in orange. The RNA is drawn as dark blue sticks. **c**, Stereo view of the SmpB–tmRNA_Δ complex. The secondary structure elements of SmpB (blue) are labelled in black. tmRNA_Δ is drawn with grey sugar-phosphate backbone and yellow bases.

dissociation of the ribosomal subunits. The released, tagged protein is recognized and degraded by various cellular proteases⁵. The tmRNA–SmpB rescue system is ubiquitous in bacteria, and is also found in some chloroplasts and mitochondria⁶.

A major component of the trans-translation system is the tRNA-like cloverleaf region (Fig. 1a)¹, which is formed by the 5' and 3' ends of the large 10S tmRNA molecule. The tRNA mimic domain of tmRNA from *Aquifex aeolicus* used in this study (tmRNA_Δ) encompasses 68 bases, including nucleotides 1 to 26 from the 5' end and 310 to 347 of the 3' end of tmRNA (Fig. 1a). The remaining bases of tmRNA were substituted by a tetraloop. Both wild-type tmRNA and tmRNA_Δ bind SmpB with a similar high affinity of ~20 nM (data not shown). Although secondary and tertiary structure models of tmRNA^{7,8} and two solution structures of SmpB^{9,10} have been reported, no three-dimensional structural information is available on the biologically active SmpB–tmRNA complex.

Here we present the 3.2-Å crystal structure of the ribonucleoprotein complex between tmRNA_Δ and SmpB. The 2F_o – 2F_c electron density map along with the final tmRNA_Δ model is presented in Fig. 1b, and Supplementary Table 1 contains statistics about diffraction data and model refinement. The overall structure shows that SmpB binds to single-stranded RNA stretches in the tmRNA_Δ elbow region, forming a one-to-one complex (Fig. 1c). There are substantial conformational differences between tmRNA_Δ and the classical tRNA^{Phe} structures reported 30 years ago¹¹, suggesting a model of tmRNA–SmpB binding to the ribosome.

The tmRNA_Δ molecule can be subdivided into four distinct domains resembling the cloverleaf-like features of tRNAs^{12,13}. The four domains include the acceptor stem as well as the TΨC-, D- and H5-arms (D-arm in tmRNA consists of only a loop). The H5-arm is joined to the TΨC-arm by a short connector loop (Fig. 1a). The tmRNA_Δ acceptor stem and the first 5 nucleotides of the D-arm are not included in the refined model, since this region of the electron density map is poorly defined (Fig. 1b). Nevertheless, the acceptor stem has been modelled as a helical extension of the TΨC-stem (Fig. 2) to facilitate the comparison between tmRNA_Δ and tRNA (Fig. 2). The tmRNA_Δ molecule adopts an open L-shaped conformation with H5 and the acceptor each forming one arm of the L. The TΨC- and D-arms meet in the elbow region of the L. In contrast to the helical stems observed in the H5- and TΨC-arms, the tmRNA_Δ D-arm is single-stranded and several bases are extended towards SmpB (Fig. 1c). This conformation is significantly different from the conformation of the D-arms of tRNAs^{11,14}. These observations confirm previous secondary structure predictions⁸, suggesting that the D-loop of tmRNA lacks secondary structural elements and is significantly shorter than its tRNA counterpart. The extended conformation of the D-arm and the reduction of tertiary interactions with the TΨC-arm probably increase the flexibility in the elbow region of tmRNA_Δ allowing for different conformational states of the acceptor stem in the crystal. The approximate value of 120° predicted from the crystal structure agrees well with a range of possible values determined in solution using transient electric birefringence experiments¹⁵. The most striking feature of the molecule is the conformation of the TΨC-arm. It is rotated by 90° around the helical axis of the acceptor stem (Fig. 2b) such that the bases of the D-loop and the short connector loop are exposed, providing a platform for interactions with SmpB (Fig. 1c).

SmpB adopts an extended oligonucleotide-binding (OB) fold¹⁰ consisting of a central β-barrel surrounded by three peripheral α-helices (Fig. 1). Strands β2, β4, β5, β7, β8 and β9 form the core of the β-barrel, which is enclosed on one side by the long helix α3 and on the other side by strands β1, β3, β6 and helix α2. No electron density was visible for the amino-terminal residues 1 and 2, the loop residues 73 to 75 and the carboxy-terminal residues 131 to 156. It was previously shown that the positively charged C terminus is disordered in solution¹⁰, and we did not observe any interactions between the tail and tmRNA domains present in our structure.

SmpB binds to the elbow region of tmRNA_Δ on the D-loop side. The contacts between SmpB and tmRNA_Δ (Fig. 3a) involve a concave protein surface on the narrow side of the OB fold, which interacts with the D- and connector-loops of tmRNA_Δ. The protein surface is formed by strands β2, β7 and β9 and the N termini of helices α2 and α3, and consists of a hydrophobic core surrounded by positively charged residues. Interestingly, SmpB uses a unique binding surface formed by elements inserted into the OB fold (α2, β2 and the loop in between) to interact with tmRNA_Δ instead of the standard OB fold binding face. 2,023 Å² of surface area are buried in the RNA–protein interface. Sequence specific interactions between SmpB and tmRNA_Δ are clustered around the 3' end of the D-loop (Fig. 3b), where highly conserved nucleotides are in close proximity to amino acids of strands β2, β7 and β9 (Fig. 3c). The last residue in the D-loop (G18) sticks into a pocket formed by both hydrophobic and charged amino acids in the short loop between strand β7 and helix α3. Conserved nucleotides are also located at the top of the H5 arm including A19 and G317, which form a non-Watson–Crick base pair as previously suggested¹⁶, but their contacts to SmpB are mediated by the RNA backbone rather than by the bases (Fig. 3c). Electrostatic interactions are observed between the positive dipole

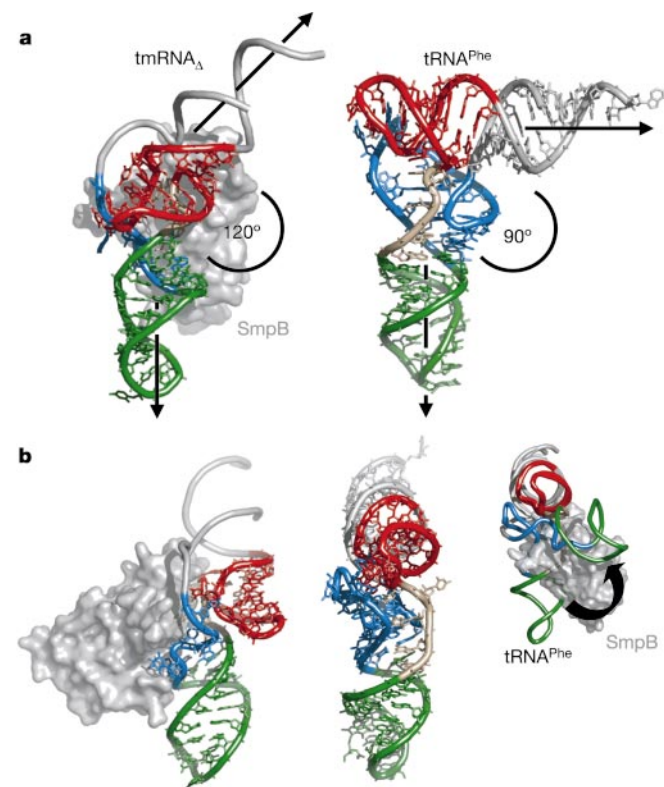
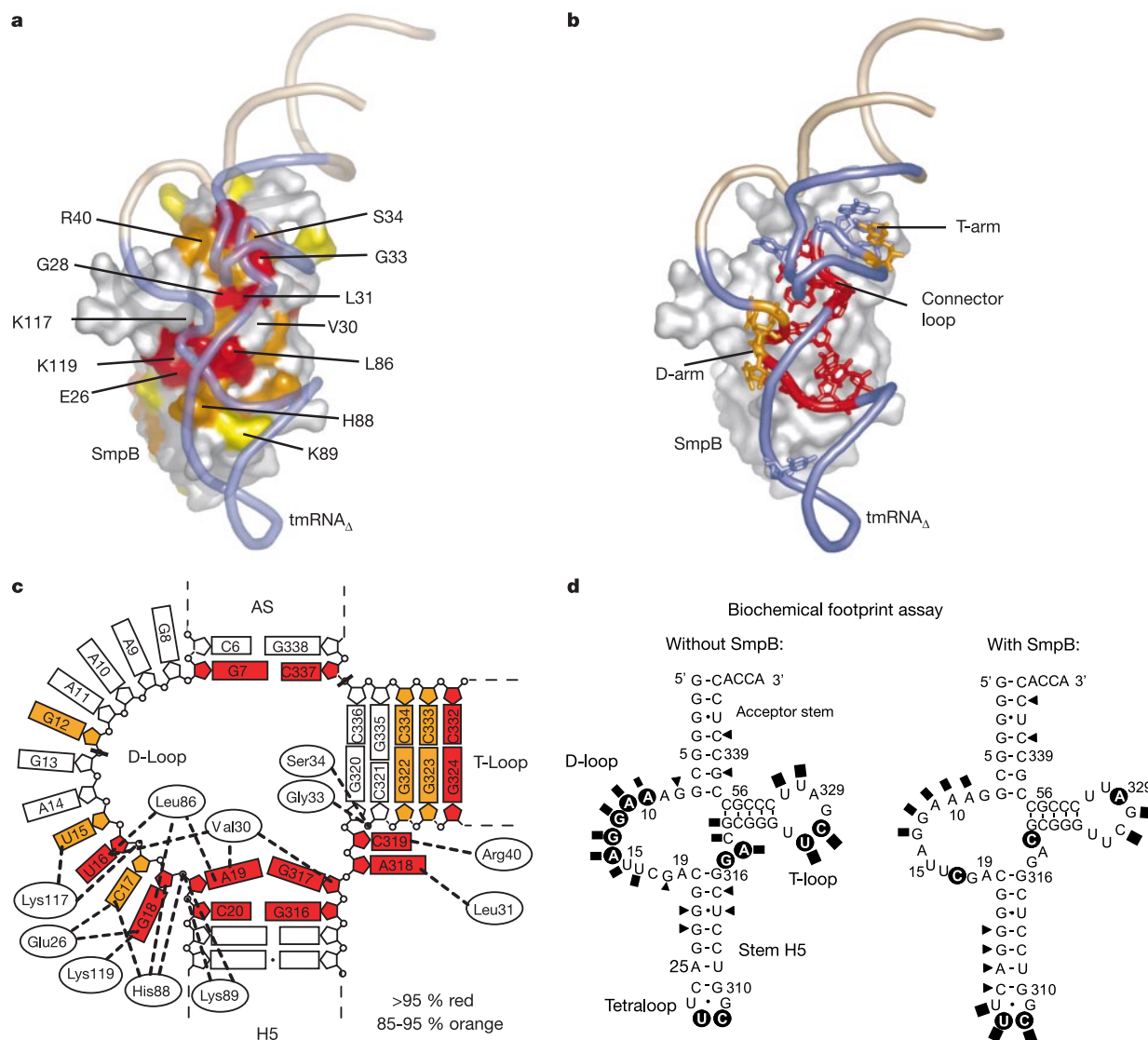


Figure 2 Two views of tmRNA_Δ in comparison with tRNA^{Phe} (ref. 14). Domains are colour-coded as in Fig. 1. SmpB is represented as semitransparent grey surface. The poorly ordered acceptor stem and five nucleotides in the D-loop have been modelled as a helical extension of the TΨC-arm into an open region of the crystal lattice with discontinuous electron density. **a**, Side views of the SmpB–tmRNA_Δ complex (left) and tRNA^{Phe} (right). Compared to Fig. 1c, the SmpB–tmRNA_Δ complex is rotated 90° counterclockwise (c.c.) around the H5 helical axis. tmRNA_Δ has a significantly larger interhelix angle (~120°) than tRNAs, but the overall L-shape of the molecules is similar. **b**, View of the back of the SmpB–tmRNA_Δ complex (left) rotated 90° c.c. around the H5 helical axis compared to **a**. The T-arm of the tmRNA is rotated approximately 90° around its helical axis, resulting in a significantly different orientation of this arm relative to the rest of the molecule than is observed for tRNA^{Phe}. Inset, tmRNA_Δ and tRNA are superimposed via their TΨC-arms (r.m.s.d. of TΨC arm phosphate atoms is 1.5 Å). Anticodon stem of tRNA^{Phe} and the H5 arm of tmRNA_Δ point in orthogonal directions. In this orientation, SmpB is partly overlapping with the tRNA^{Phe} acceptor stem (see also Fig. 4).

of helices $\alpha 2$ and $\alpha 3$ and the phosphates of C321 and A19. The SmpB amino acids involved in tmRNA binding are highly conserved in bacteria (Fig. 3a).

Enzymatic and chemical footprints of tmRNA $_{\Delta}$ in complex with SmpB also indicate that the protein interacts with both D- and connector loops of tmRNA (Fig. 3d). Furthermore, many secondary and tertiary RNA interactions observed in the crystal structure are supported by the biochemical probing experiments in solution. Nevertheless, unbound tmRNA $_{\Delta}$ in solution compared to the tmRNA $_{\Delta}$ in complex with SmpB has slightly different conformation in the acceptor stem, H5 and T loop regions, as judged by the difference in accessibility of these domains to lead acetate and RNases S₁ and V₁. These experiments suggest that SmpB binding limits the range of conformations that tmRNA $_{\Delta}$ can adopt.

Trans-translation requires tmRNA to be charged with alanine by AlaRS before delivery to the stalled ribosome. In contrast to other aminoacyl-tRNA synthetases (aaRS)¹⁷, AlaRS primarily relies on simple conserved features within the acceptor stem to recognize its substrate tRNA^{Ala} (refs 18, 19), which may be the reason why all tmRNAs are charged with alanine^{12,13}. AlaRS charges both wild-type tmRNA and tmRNA $_{\Delta}$ with similar efficiency *in vitro* (results not shown), indicating that our truncated unmodified construct, tmRNA $_{\Delta}$, is biologically active as a tRNA. Although no structural information is available on AlaRS, sequence similarities indicate that AlaRS is a class IIc aaRS, which approach the 3' ACCA end of tRNAs from the major groove side of the acceptor stem¹⁷. Since SmpB binds to the opposite side of the tmRNA, complex formation does not interfere with tmRNA aminoacylation. After aminoacyla-



tion EF-Tu is required to deliver the SmpB–tmRNA_Δ complex to stalled ribosomes. Consistent with previous biochemical studies²⁰, there are no overlaps between the SmpB binding site and the canonical site of EF-Tu interaction, as predicted by superimposing the TΨC-arms of tmRNA_Δ and tRNA.

An intriguing question is how stalled ribosomes are recognized by the trans-translation machinery. After delivery to stalled ribosomes, the aminoacylated SmpB–tmRNA complex might be accommodated into the ribosomal A-site via GTP hydrolysis by EF-Tu, as for canonical tRNAs, which depends on the correct pairing of mRNA codons with cognate tRNA anticodons.

The structure of the SmpB–tmRNA_Δ complex fitted onto the A-site bound tRNA shows that, owing to the 90° rotated TΨC-arm, SmpB is oriented towards the decoding centre of the small ribosomal subunit (Fig. 4). The H5-stem and the remaining tmRNA extend towards the elongation-factor binding region underneath the L7/L12 stalk of the large ribosomal subunit. Although SmpB does not directly interact with the decoding centre in our model, its proximity suggests that the interactions could be mediated by the

positively charged C terminus of SmpB acting as an anticodon stem mimic.

Steric clashes of SmpB with the P-site tRNA or of the long disrupted H5 helix with the decoding centre are prevented by the rotation of the TΨC-arm around the acceptor stem axis placing the bulk of the tmRNA away from the intersubunit space. This would allow tmRNA to loop outside the ribosome near the elongation-factor binding side, and deliver its mRNA part into the active site of the small ribosomal subunit. This is a reasonable starting point for the progression of the large tmRNA through the ribosomal active site, preventing encircling or steric clashes with various structural features of the ribosome such as the L7/L12 stalk or the 30S neck region.

It is not clear how the SmpB–tmRNA_Δ complex would fit into the P-site of the ribosome after translocation, since it is much larger than a tRNA molecule. In particular, binding of the next tRNA to the ribosome requires an accessible A-site. It is possible that SmpB transiently dissociates at this stage or that the tmRNA possesses sufficient flexibility in the elbow region to swing SmpB and the H5 arm completely out of the way.

We have been made aware of a paper on the cryo-EM reconstruction of the kirromycin stalled tmRNA–EF-Tu complex with the ribosome²¹. Although there is a difference in the orientation of the SmpB and tmRNA H5 helix when the EM structure is compared to our model, it probably reflects a different stage of tmRNA accommodation into the ribosomal active site. The EM reconstruction captures tmRNA before accommodation of its acceptor stem into the peptidyl transferase centre, whereas our model places the X-ray structure into the ribosomal A-site after accommodation. □

Methods

Sample preparation and crystallization

A. aeolicus tmRNA_Δ was *in vitro* transcribed using T7 RNA polymerase and purified by denaturing polyacrylamide gel electrophoresis (PAGE)²². The SmpB gene was PCR amplified from *A. aeolicus* genomic DNA (a gift from K. Stetter and R. Huber, Universität Regensburg, Germany), cloned into pET21a and expressed in *Escherichia coli* BL21(DE3)/pLysS (Novagen) at 37 °C after induction with 1 mM IPTG. SmpB was purified by metal affinity purification with NiNTA resin (Qiagen) and cation exchange chromatography using a Source15S column (Pharmacia). SmpB and tmRNA_Δ were mixed in presence of 10 mM magnesium chloride and purified by anion exchange chromatography.

Crystals grew to a maximum size of 300 × 200 × 200 μm³ in 10 days after mixing 6 mg ml^{−1} SmpB–tmRNA_Δ complex with 10% 2-propanol, 100 mM magnesium acetate, 15 mM ammonium acetate, 50 mM sodium cacodylate pH 6.8, 0.2 mM gentamicin or neomycin B²³, and 16 mM CaCl₂ at 19 °C using sitting drops. Analysis of dissolved crystals by SDS–PAGE showed that both RNA and protein molecules were intact. Prior to flash freezing in liquid propane, the crystals were cryo-protected in 25% MPD. For multi wavelength anomalous dispersion (MAD) experiments, crystals were transferred to a drop containing 20 mM pentaamine (trifluoromethanesulphonate) osmium (III) trifluoromethane for 2 h before freezing. Crystals belonged to the space group P4₃2₁2 (*a* = *b* = 99.7 Å, *c* = 207.1 Å) with two SmpB–tmRNA_Δ molecules in the asymmetric unit (solvent content of 56%).

Structure determination and refinement

Native data were collected to 3.2 Å (*R*_{merge} = 7.5% with overall completeness 98.4%) (Supplementary Table 1). For structure solution, MAD data were collected at the peak of the osmium LIII-edge, the inflection and a remote wavelength. Both native and MAD data were collected at 100 K on beamline X06SA at the Swiss Light Source (SLS, PSI Villigen, Switzerland) and processed with the HKL package²⁴. Four osmium sites were identified with SOLVE²⁵ and refined in CNS²⁶ (overall figure of merit, 0.43).

Initial MAD phases were improved by solvent-flipping and averaging using native data to 3.2-Å resolution. Based on these maps, an initial structure was built using O²⁷ and improved by iterative cycles of model building into composite omit maps and refinement. The RNA and protein backbones were defined by continuous stretches of electron density. Bulky protein side chains were visible in our maps and served as reference positions for SmpB structure building. In case of RNA, the structure was built based on initial positioning of the tetraloop into electron density maps. From this starting point, the register of the sequence could be preserved, although at a resolution of 3.2 Å discrimination between purines and pyrimidines was not possible. In the crystal, two SmpB–tmRNA_Δ molecules were related by 179° non-crystallographic, pseudo-twofold symmetry. While one RNA molecule in the asymmetric unit was well characterized by electron density, equivalent parts of the second RNA molecule suffered from discontinuous electron density. The refined structure comprises 45 residues of the better-defined RNA molecule and 24 residues of the other RNA molecule in the asymmetric unit.

Our results presented here are based on the well-defined RNA molecule of the asymmetric unit. Both protein molecules in the asymmetric unit were equally well

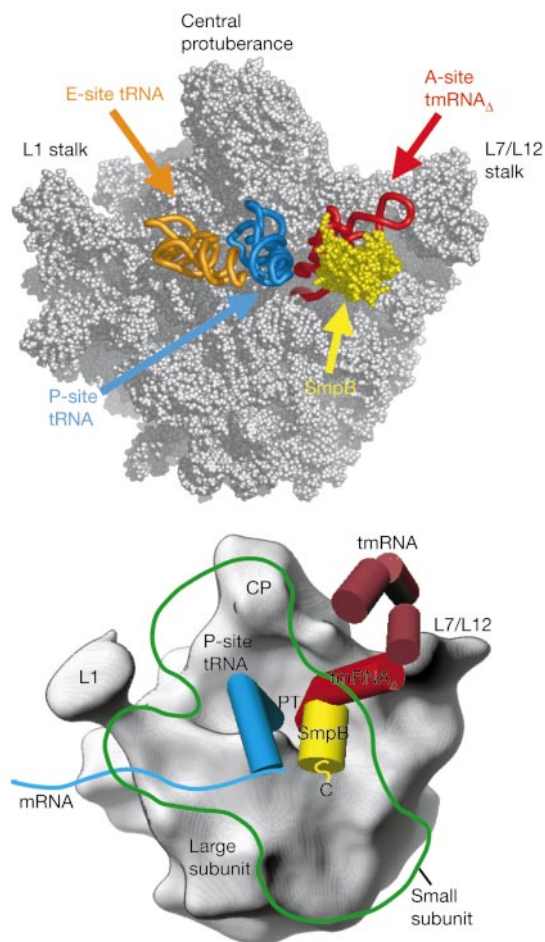


Figure 4 Model of the interactions between the SmpB–tmRNA_Δ complex and the ribosome. Top, the SmpB (yellow)–tmRNA_Δ (red) complex is fitted into the peptidyl transferase centre of the 50S ribosomal subunit (grey)²⁰. P- and E-site tRNAs are coloured blue and orange, respectively. The structure of the tmRNA_Δ complex was positioned by superimposing its TΨC-arm onto the equivalent region of the tRNA (see Fig. 2b inset) bound to the A-site of the ribosome³⁰. Bottom, scheme of the SmpB–tmRNA complex bound to the 50S subunit. SmpB (yellow) is oriented towards the active site on the 30S subunit while H5 of the tmRNA_Δ complex extends towards the elongation-factor binding region. Additional tmRNA domains (cylinders in dark red) wrap around the 30S subunit indicated with green outline²¹.

defined, and residues 3 to 130 of each SmpB molecule were included into refinement. The structure was refined to 3.2 Å using simulated annealing in CNS and had an R_{free} value of 36.0% and an R_{cryst} of 30.2% with an overall B-value of 108.4 Å². The r.m.s.d. values in bond length were 0.007 Å and in bond angles 1.5°. Figures were computed with Pymol²⁸.

Received 20 March; accepted 23 May 2003; doi:10.1038/nature01831.

- Karzai, A. W., Roche, E. D. & Sauer, R. T. The SsrA-SmpB system for protein tagging, directed degradation and ribosome rescue. *Nature Struct. Biol.* **7**, 449–455 (2000).
- Keiler, K. C., Waller, P. R. & Sauer, R. T. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science* **271**, 990–993 (1996).
- Karzai, A. W., Susskind, M. M. & Sauer, R. T. SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA). *EMBO J.* **18**, 3793–3799 (1999).
- Rudinger-Thirion, J., Giege, R. & Felden, B. Aminoacylated tmRNA from *Escherichia coli* interacts with prokaryotic elongation factor Tu. *RNA* **5**, 989–992 (1999).
- Gottesman, S., Roche, E., Zhou, Y. & Sauer, R. T. The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev.* **12**, 1338–1347 (1998).
- Keiler, K. C., Shapiro, L. & Williams, K. P. tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. *Proc. Natl Acad. Sci. USA* **97**, 7778–7783 (2000).
- Williams, K. P. & Bartel, D. P. Phylogenetic analysis of tmRNA secondary structure. *RNA* **2**, 1306–1310 (1996).
- Felden, B. *et al.* Probing the structure of the *Escherichia coli* 10Sa RNA (tmRNA). *RNA* **3**, 89–103 (1997).
- Someya, T. *et al.* Solution structure of a tmRNA-binding protein, SmpB, from *Thermus thermophilus*. *FEBS Lett.* **535**, 94–100 (2003).
- Dong, G., Nowakowski, J. & Hoffman, D. W. Structure of small protein B: The protein component of the tmRNA-SmpB system for ribosome rescue. *EMBO J.* **21**, 1845–1854 (2002).
- Kim, S. H. Three-dimensional structure of yeast phenylalanine transfer RNA: Folding of the polynucleotide chain. *Science* **179**, 285–288 (1973).
- Ushida, C., Himeno, H., Watanabe, T. & Muto, A. tRNA-like structures in 10Sa RNAs of *Mycoplasma capricolum* and *Bacillus subtilis*. *Nucleic Acids Res.* **22**, 3392–3396 (1994).
- Komine, Y., Kitabatake, M. & Inokuchi, H. 10Sa RNA is associated with 70S ribosome particles in *Escherichia coli*. *J. Biochem. (Tokyo)* **119**, 463–467 (1996).
- Shi, H. & Moore, P. B. The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA* **6**, 1091–1105 (2000).
- Stagg, S. M., Frazer-Abel, A. A., Hagerman, P. J. & Harvey, S. C. Structural studies of the tRNA domain of tmRNA. *J. Mol. Biol.* **309**, 727–735 (2001).
- Hanawa-Suetsugu, K., Bordeau, V., Himeno, H., Muto, A. & Felden, B. Importance of the conserved nucleotides around the tRNA-like structure of *Escherichia coli* transfer-messenger RNA for protein tagging. *Nucleic Acids Res.* **29**, 4663–4673 (2001).
- Ibba, M. & Soll, D. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69**, 617–650 (2000).
- Hou, Y. M. & Schimmel, P. A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* **333**, 140–145 (1988).
- McClain, W. H., Foss, K., Jenkins, R. A. & Schneider, J. Four sites in the acceptor helix and one site in the variable pocket of tRNA(Ala) determine the molecule's acceptor identity. *Proc. Natl Acad. Sci. USA* **88**, 9272–9276 (1991).
- Barends, S., Karzai, A. W., Sauer, R. T., Wower, J. & Kraal, B. Simultaneous and functional binding of SmpB and EF-Tu-TP to the alanyl acceptor arm of tmRNA. *J. Mol. Biol.* **314**, 9–21 (2001).
- Valle, M. *et al.* Visualizing tmRNA entry into a stalled ribosome. *Science* **300**, 127–130 (2003).
- Price, S. R., Ito, N., Oubridge, C., Avis, J. M. & Nagai, K. Crystallization of RNA-protein complexes. I. Methods for the large-scale preparation of RNA suitable for crystallographic studies. *J. Mol. Biol.* **249**, 398–408 (1995).
- Corvaisier, S., Bordeau, V. & Felden, B. Inhibition of transfer messenger RNA aminoacylation and trans-translation by aminoglycoside antibiotics. *J. Biol. Chem.* **278**, 14788–14797 (2003).
- Otwinowski, Z. & Minor, W. Processing X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Terwilliger, T. C. Automated structure solution, density modification and model building. *Acta Crystallogr. D* **58**, 1937–1940 (2002).
- Brunker, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- DeLano, W. L. The PyMOL Molecular Graphics System (<http://www.pymol.org>) (2002).
- Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
- Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Schulze-Bries, T. Tomizaki and A. Wagner (SLS, Villigen) D. Sargent, and the team at Swiss Norwegian Beamline (ESRF, Grenoble) for assistance in data collection, V. Ramakrishnan for comments on the manuscript, and D. Goven for help in the footprint assays. P.W.H. is supported by an EMBO fellowship. This work was supported by the Swiss National Science Foundation (SNSF), the NCCR Structural Biology programme of the SNSF, and a Young Investigator grant from the Human Frontier Science Program.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.B. (ban@mol.biol.ethz.ch). Structure coordinates have been deposited in the Protein Data Bank under accession code 1P6V.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Global Head of Naturejobs:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/

Iberia/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@nature.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: r.asami@nature.com

naturejobs

Physical paradoxes

During a recent visit to some federal physics labs near Chicago, I saw three career trends that seemed to be contradictory. First was the discovery that physics graduate students and postdocs are likely to find it hard to secure a long-term faculty post. Yildirim Mutaf, a graduate student doing his thesis research at Fermilab, notes that only about 10% of his peers would secure such positions.

Compounding matters was the sense that the search for qualified people to staff the labs is not easy. "By and large I have to look outside the United States," says Robert Rosner, chief scientist at Argonne National Laboratory.

Perhaps most perplexing of all is Fermilab's seeming insistence on contributing to its own apparent obsolescence. Several groups there are creating instruments for the Large Hadron Collider (LHC) in Switzerland, which will be seven times more powerful than Fermilab's own Tevatron collider. An operational LHC could well move the epicentre of high-energy physics away from Chicago.

How are the national labs dealing with these apparent inconsistencies? By selecting physicists who are adaptable, as well as creating new opportunities, says Rosner. Mutaf, for example, is keeping his options open, training himself in basic research techniques ready for an academic job, but making himself aware of opportunities in the private sector, such as finance. Rosner, meanwhile, is trying to attract more top physicists by sponsoring the development of new technologies and by touting generous postdoc programmes. And Fermilab scientists, by helping the LHC, are setting themselves up as collaborators.

By employing such agility, scientists at Argonne and Fermilab have managed to make three sets of opposing perceptions equally true — maybe Illinois should be renamed the quantum state.

Paul Smaglik

Naturejobs editor



Contents

REGIONS

Chicago's physical
stature

p706

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

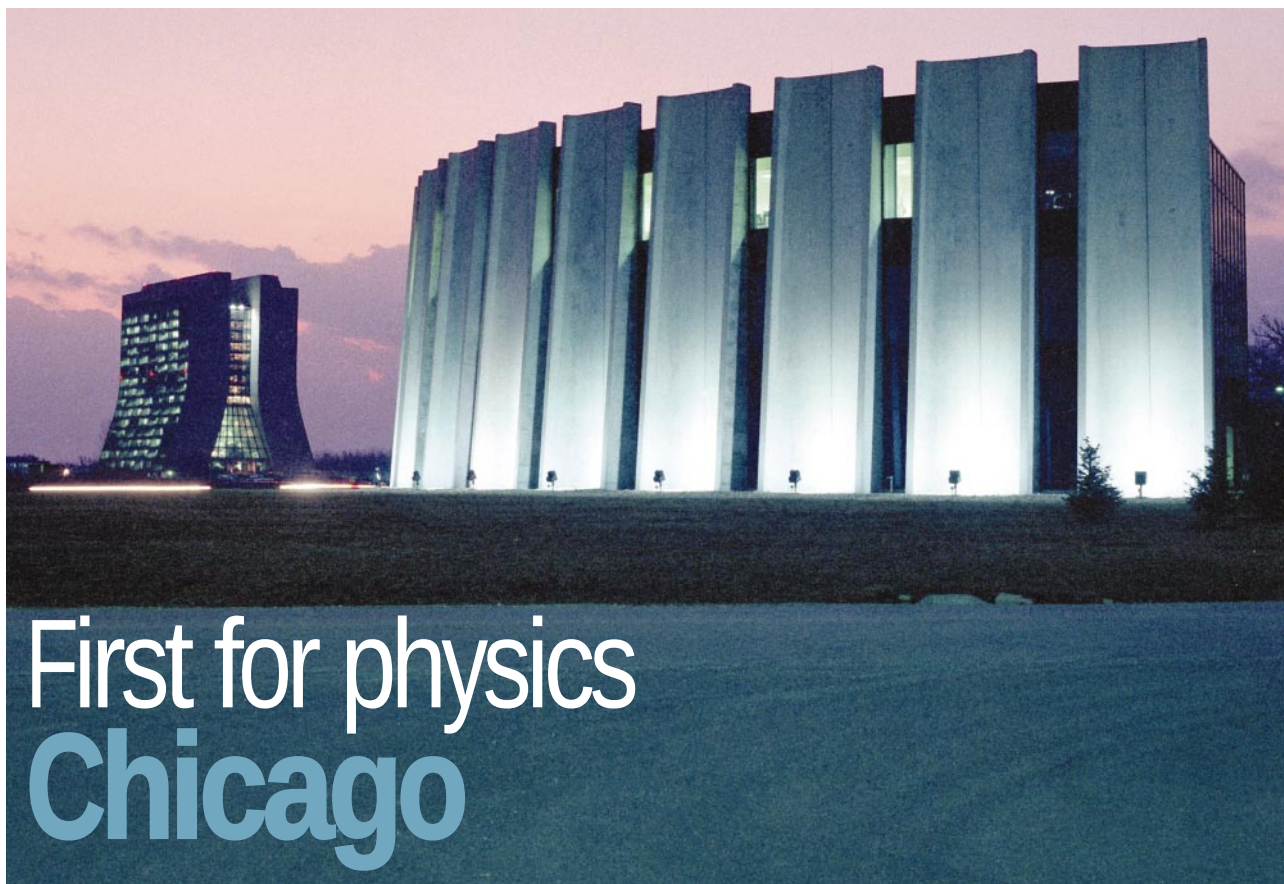
FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS



First for physics Chicago

Light fantastic: Fermilab's central lab building, Wilson Hall (left), and the Feynman computer centre.

Chicago is known as the United States' second city of comedy. But with two US Department of Energy laboratories within a 30-minute drive of each other, it may be the first city of physics. As so often in physics, this claim is underpinned by technology. The Fermi National Accelerator Laboratory (Fermilab) boasts the Tevatron, the world's most powerful particle accelerator: a 6.4-km-long ring that accelerates subatomic particles towards each other, smashes them, and then records the aftermath. Scientists hope that this will help them to understand the fundamental nature of matter and the origins of the Universe. The jewel in the crown at the Argonne National Laboratory is the Advanced Photon Source (APS), the most powerful synchrotron radiation facility in the United States. It sends high-powered beams of light into proteins or material, helping scientists to work out their structure on the atomic scale.

The two facilities employ a few thousand people, with a few thousand more visiting every year to do research. But developments in Europe and Japan could change the area's employment dynamics. CERN, the European particle-physics lab near Geneva, is building the Large Hadron Collider (LHC). When it goes online in 2007, this will be seven times more powerful than the Tevatron. Some synchrotrons in Japan and Europe are already more powerful than the APS.

The Chicago scientists are meeting the challenge. They have begun designing

technology to meet needs that haven't been addressed and are finding new uses for existing equipment. John Cooper, head of Fermilab's particle-physics division, notes, somewhat wryly, that some scientists at Fermilab are adding to the facility's obsolescence.

THE SPIRIT OF PHYSICS

Fermilab scientists are developing and building superconducting magnets for the LHC that steer particles towards a collision, and are helping to build detectors that record the paths of the fragments after the particles smash.

The spirit of physics sets the overall goal of the field above any competition, says Cooper. During the past few years, while CERN has been running at reduced capacity in order to build the LHC, many European collaborators

have sent their students to Fermilab (see *Nature* **409**, 754–755; 2001). “They need a place for their students to get trained while they are waiting for the LHC to get turned on,” Cooper says.

Will the flow of young physicists go in the other direction when the LHC goes online? “We’re not really sure,” says Cooper. What he does know is that the collaborations being formed now improve Fermilab's chances of keeping a role after it is overtaken.

Its neighbour is helping. Argonne computer scientist Ian Foster and collaborators have developed the Grid, a form of computing that makes better



Nigel Lockyer: predicts a rise the amount of data analysis.

use of existing networks and allows users in different locations to work on the same set of data. Fermilab scientists could use the Grid's powerful capacity to process and analyse a torrent of data from the LHC's detectors. Continuing to build and refine its infrastructure and software could also keep scientists at both labs busy.

Nigel Lockyer, a University of Pennsylvania professor who heads experiments at one of the Tevatron's main detectors, says that the 600 or so scientists there mainly do data analysis. The LHC's greater power and capacity will mean even more work for them.

"There will be a big effort to keep things going," says Lockyer. And with the Grid, it shouldn't much matter if the scientists are in Geneva or Illinois.

GOODBYE TO THE OWL SHIFT

Such an arrangement could save on travel between the European and US labs, says Cooper. "You could control your experiment at CERN while sitting at Fermilab," he says. It would also save visiting scientists at Fermilab having to work the 'owl shift' and live in trailers.

Fermilab's accelerator may still have a role, even once the LHC is online. Some experiments, such as those involving neutrinos, require less power. Fermilab has a small experiment, the Mini Booster Neutrino Experiment, under way, and plans to start a larger one, the Main Injector Neutrino Oscillation Search, in early 2005. The Tevatron may go on serving the neutrino field even if it is superseded for other types of experiment.

Argonne's facilities are less likely to become outdated, as the lab doesn't depend on one instrument. Even if other synchrotrons are stronger than the APS, the demand to probe protein structures and new kinds of material should keep the facility busy for quite some time. And it has room for more: its 40 beamlines, split between biology and materials research, can be increased to run 30 more experiments.

Robert Rosner, Argonne's chief scientist, has ambitions for the lab beyond the APS. A \$36-million Center for Nanoscale Materials now being built there will have an operating budget of between \$15 million and \$20 million.

Another project will expand Argonne beyond its current campus. There is already an energy-research centre on neighbouring land owned by the local airport, which is being developed for more applied research (see 'Looking for company', right). The facility will host research on fuel cells, converting hydrocarbons into hydrogen and storing and transporting the gas, says Harvey Drucker, associate laboratory director for energy and environmental research. Once the facility is complete, the lab's staff of 100 will grow to 150, with a doubling of graduate students and postdocs.

One of the biggest hopes for Argonne, in terms of both science and employment, is the Rare Isotope Accelerator (RIA), an instrument that would allow physicists to probe the interactions between particles that make up an atom's nucleus, and to make isotopes for applied-science purposes.

Such isotopes could be used to analyse properties of new materials. Engineers can understand piston wear



Robert Rosner: wants to see Argonne as a centre for nanotechnology.

by irradiating the piston wall and then measuring radiation. Materials scientists could use rare isotopes produced by the RIA to do the same with, say, artificial knee joints.

But although the Department of Energy and the National Science Foundation have called the RIA their "highest priority for major new construction", the money has not yet been committed. And Argonne is competing with Michigan State University for the \$65-million facility. There's a lot at stake. The RIA would create jobs for some 400 operators and bring in about 800 users a year by 2010,

the earliest it could be fully online, says Donald Geesaman, director of Argonne's physics division.

New facilities will play an important role in keeping the area robust, so scientists from both labs and local universities have begun banding together. Argonne's Kwang-Je Kim is leading a multi-institution effort to set up the Institute for Advanced Accelerator Physics, where scientists will ponder how instruments such as the APS could be tweaked into a more powerful light source, and consider technologies not yet in use for new kinds of accelerator.

STRENGTH IN NUMBERS

The Illinois Consortium for Accelerator Research, a group of universities and institutes formed to back Fermilab, has allowed local universities to pool their resources with the two national labs, resulting in more state funding for research salaries.

Consortium director Chris White, assistant professor of physics at the Illinois Institute of Technology, was drawn to the area by such interactions. He participates in research projects at both federal labs. And when the LHC gets turned on? "That's many years out," says White. Meanwhile, "there's no better place than Chicago to get a job in physics". ■

Paul Smaglik is editor of *Naturejobs*.

Looking for company

Before the telecom bust a few years ago, local powerhouses such as Lucent and Motorola supplied many postdocs and students who did stints at Fermilab or the Argonne National Laboratory with more permanent employment. The DuPage Airport Authority may provide a new engine for opportunity, by turning one of the few large tracts of undeveloped land near Chicago into a science and technology park. Argonne is the first

to commit to building applied research facilities on the site, 324 hectares between the airport and Fermilab. New jobs are expected to arise as companies emerge from the energy-research effort, and as materials-science research is increased at university and federal labs in the area. The state of Illinois has pledged \$34 million to develop the DuPage Technology Park's infrastructure, now under construction. P.S. www.dupagetechnology.com